ESSAYS IN LABOR ECONOMICS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Lingwen Zheng

August 2014

ESSAYS IN LABOR ECONOMICS

Lingwen Zheng, Ph.D.

Cornell University 2014

The first chapter of this dissertation examines the phenomenon of labor market segregation. Using a regression discontinuity (RD) design, I exploit the variation in base-year minority shares across single-establishment firms to document the dynamics of establishment-level segregation in two five-year intervals: 1995-2000 and 2000-2005. Using the Longitudinal Employer-Household Dynamics (LEHD) infrastructure files, I first show that systematic establishment-level segregation still exists in all industries. Then, I show that the dynamics of segregation among these single-establishment firms are non-linear and exhibit "tipping" patterns in both five-year intervals, although the magnitude is much larger in the earlier time period. The observed tipping pattern is primarily driven by non-Hispanic whites leaving. The effect due to minorities entering is much smaller. Alternative explanations such as non-linear changes in establishment characteristics or omitted variables do not explain the observed changes in minority shares. Finally, I find that, unlike the 1995-2000 period, during which tipping behavior seems to have been driven equally by blacks and Hispanics, Hispanics are the sole driving force in the 2000-2005 period. Overall, this chapter provides the first suggestive evidence that the dynamics of establishment-level segregation are highly nonlinear and exhibit a tipping pattern.

The second chapter of the dissertation describes the technical linking process and examines the properties and the qualities of the crosswalk files. The crosswalk between the Longitudinal Employer-Household Dynamics (LEHD) infrastructure file system and the Census Business Register (BR) is authorized as part of the LEHD Infrastructure Project. This document describes the LEHD - BR crosswalk and its component inputs: the Business Register, Longitudinal Business Database (LBD), and the LEHD Infrastructure File

system. The output files include the LEHD - BR crosswalk at both the establishment and employer levels. These output files can facilitate linking a wide range of contextual variables relating to characteristics of the current and prior employers and co-workers of current employees. Match and non-match rates for various populations are defined and estimated in order to examine the properties and quality of the LEHD - BR crosswalk output files.

The third chapter of this dissertation exploits plausibly exogenous changes in family size caused by the initial implementation and subsequent relaxations in China's One Child Policy to estimate the causal effect of family size on educational attainment. I find that the average family size has decreased substantially since the One Child Policy implementation. By employing an Instrumental Variable estimation strategy, I find clear evidence indicating that there is indeed a negative trade-off between child's quantity and quality in urban China. An additional child can lead to a decrease of 1.2 years of schooling. A simple back-of-the-envelope calculation reveals that the implementation of the One Child Policy has significantly increased the average completed years of schooling by approximately 0.68 years in urban China. This effect is in fact larger for women than for men. No negative trade-off effect is found for the rural households in the sample.

BIOGRAPHICAL SKETCH

Lingwen Zheng was born and raised in Beijing, China. She graduated from Beijing Huiwen High School in 2001. She earned her B.E. in Urban and Regional Planning and B.A. in Economics from Peking University in 2006. She came to the U.S. in 2006, where she earned her M.R.P. in City and Regional Planning from Cornell University in 2008, her M.A. in Economics from Cornell University in 2013, and expects to receive her Ph.D. in Economics from Cornell University in August 2014. During her last two years at Cornell, she worked to complete her dissertation using the U.S. Census Bureau's confidential data while working for Professor John M. Abowd as his research assistant.

To my wonderful parents Peishan Zheng and Fuying Wang.

To my beloved husband Andrew Slocum.

To my forever companion Mimi.

献给给予我支持和关爱的父母，丈夫，以及咪咪

ACKNOWLEDGEMENTS

I would like to express my gratitude to the members of my thesis committee: John M. Abowd, Francine D. Blau and Jordan D. Matsudaira. I especially appreciate the guidance and mentoring that John and Fran have provided me during my five years at Cornell University. As for John, I am eternally grateful for the opportunity to work as your research assistant and I have genuinely enjoyed working with you. Your mastery of frontier economic research and work ethnics has benefited me tremendously. You have also made me challenge myself in ways I never thought I could. I will benefit from everything I have learned while I was working for you throughout my professional career. As for Fran, your rigorous research styles in addition to your mastery of frontier economic research have been a major source of motivation for me. Your work ethics, strive for perfection, and genuine care for your students will continue motivating me to be my best both professionally and personally.

I am also indebted to my colleagues at the Labor Dynamics Institute whom I have learned a great deal from directly and indirectly. From Lars Vilhuber, I have learned how to program in SAS using the most sophisticated and efficient methods. I have benefited greatly from your knowledge, thoroughness, skills, and quality work. I greatly appreciate all the help you have provided me throughout the years. Nellie Zhao, my colleague and friend, is who has given me immense support during my last year while I was preparing for and engaged in the formal job market for Economists. She is also someone I have learned a great deal about Economics and Economic research. As for the entire Labor Dynamics Institute research group, it has been a great pleasure to work with you all.

I would like to express my gratitude to Professor Mildred Warner in the Department of City and Regional Planning at Cornell University. I am extremely grateful for your mentoring throughout the years. Your confidence in me and your faith in me as a student and as a researcher has given me strength during the last few years. I will benefit a great deal from all the professional advice you have given me during my professional career. I

am extremely grateful to have known you and your family. I am lucky to have you as my advisor, mentor, and friend.

I am grateful for the love and support of my parents throughout this process. To my mom, who has always been by my side and encouraged me to reach higher, you continue to inspire me to press forward. To my father, who has always worked hard himself, thank you for teaching me the meaning and importance of diligence and perseverance. You both are the ultimate inspiration for me to be my best.

Last but not the least, I thank Andrew, my husband. I cannot imagine having gone through these past several years without having you in my life. With your big heart, patience, understanding, unconditional love, and support, you have always manage to keep me in balance. I am eternally grateful to have you as my husband.

Contents

List of Tables

List of Figures

Chapter 1

Tipping Points: The Dynamics of Workplace Segregation by Race and Ethnicity

1.1 Introduction

With the rise in minority shares in the U.S., research has demonstrated that racial and ethnic segregation still prevails in residential places (Cutler et al., 1999; Ananat, 2011; Card et al., 2008a), in schools (Caetano and Maheshri, 2014), and in the labor market (Higgs, 1977; Albelda, 1986; Carrington and Troske, 1998; Hellerstein and Neumark, 2003; Hellerstein et al., 2008; Gradín et al., 2011). Segregation in the labor market is said to exist if members of different groups are more likely to work with coworkers who are more like themselves than would be predicted by a random allocation of workers to firms (Hellerstein and Neumark, 2008).

Labor market segregation by race and ethnicity is an important area of research because, leaving aside the potential social issues, it may account for - at least in a statistical sense - a significant share of wage differentials between whites and various minority groups (Hellerstein

et al., 2008). To date, most empirical research has documented the magnitude of segregation at the industry-level or at the occupation-level (Higgs, 1977; Albelda, 1986; King; Gradín et al., 2011), and identified some of its possible causes (Carrington and Troske, 1998; Hellerstein and Neumark, 2003, 2008). Segregation at the workplace level, however, has been noticeably under-studied. Data constraints, in particular, the lack of matched employer-employee data, have been a major cause. Nonetheless, research on workplace segregation should be emphasized because it may be much more salient for interactions between racial and ethnic groups than is residential segregation. In fact, Hellerstein et al. (2008) found that racial and ethnic segregation at the three-digit industry level is usually one-third as large as the establishment-level segregation experienced by minority workers. In this paper, I will first document the extent of racial and ethnic segregation at the workplace level using the Longitudinal Employer-Household Dynamics (LEHD) infrastructure files – a matched employer-employee dataset.

In the literature on residential segregation, Card et al. (2008a) have shown that once the base-period minority share in a census tract reaches a certain level, white flight occurs. They define such phenomenon as the evidence of tipping and assert that tipping process can capture the underlying mechanism that leads to residential segregation. Caetano and Maheshri (2014) have demonstrated that a similar tipping effect also exist in school segregation. The dynamic process of labor market segregation, on the other hand, is not as well understood. The dynamics of labor market segregation have crucial implications for understanding its persistence. Better understanding of the dynamics might also facilitate the evaluation of policy measures aimed at promoting racial and ethnic integration in the context of the labor market. The second goal of this paper is to use the LEHD infrastructure files to begin to unravel the dynamics of workplace segregation by race and ethnicity.

Figure 1.1 illustrates that the sudden percentage changes in net establishment-level white employment, defined as the percentage change in white employment net of the percentage

change in minority employment, [1] in all industries pooled, and in the service-producing NAICS supersector [2] appear to be related to a workplace's base-period minority share. Here and throughout the paper, minorities are defined as nonwhites and white Hispanics; whites are defined as non-Hispanic whites only. Each plot depicts the mean percentage changes in net white employment from 1995 to 2000 deviated from the average of the same variable within the NAICS sector, grouping establishments into one-percentage-point wide cells by the minority share in 1995. Figure 1.1 shows striking evidence of non-linearities in the percentage change in net white employment. Such non-linearities may be a function of base-period minority share. This is suggestive of the existence of a "tipping phenomenon" at the workplace, where workplace minority composition increases rapidly once the the base-year minority share reaches or exceeds a critical threshold. The threshold level at which this rapid change occurs is called a "tipping point (Card et al., 2008a; Pan, 2010)."

What theoretical model can explain these non-linear patterns of workplace minority composition changes? I hypothesize that the classic social interaction model posited by Schelling (1971) can account for this empirical finding. A large body of work has focused on theorizing about the causes of segregation, for instance, the statistical discrimination models (Phelps, 1972; Arrow, 1973), the taste-based discrimination theory (Becker, 1971; Blau et al., 2010), the "pollution" theory of discrimination (Goldin, 2002), and other models using supply and demand in the labor market (Altonji and Blank, 1999; Kaufman, 2002; Reskin et al., 1999; Sorensen, 2004). However, these explanations overlook the possible effect of "post-hiring" dynamics on workplace composition (Sorensen, 2004) and provide little insight on the underlying mechanisms driving the segregation. Schelling (1971), on the other hand, developed the social interaction model to show that substantial segregation can arise from social interactions and weak prejudice against one group (Card et al., 2008a; Pan, 2010). Since its

[1] The percentage change in establishment-level white employment is expressed as the change in white employment as a percentage of the total employment in a single-establishment firm in the base year. The percentage change in establishment-level minority employment is defined in a similar manner.

[2] Refer to Appendix 1.7.A. for definitions. In this paper, industries and NAICS sectors are used interchangeably.

development, Schelling's model has been used in many areas of research such as residential segregation (Card et al., 2008a) and gender segregation in the labor market (Pan, 2010).

This paper studies the possible effects of "post-hire" dynamics on workplace composition. It attempts to unravel the underlying dynamics of workplace segregation by race and ethnicity. I test whether establishments exhibit "tipping"-like behavior in response to firm-specific shocks in minority labor supply that occur over two five-year intervals: 1995-2000 and 2000-2005. I also analyze the shifting composition of single-establishment firms in the U.S. labor market, which could help explain the persistence of segregation and shed light on the potential effectiveness of policies promoting workplace integration. Only Sorensen (2004) has investigated workplace dynamics and minorities in the paper modeling the relationship between worker turnover and the racial composition of the employing establishment's workforce using a three-year panel data of one multi-unit firm. The author finds that the worker turnover rate is negatively correlated with the minority share in that firm.

This paper uses a Regression Discontinuity (RD)-tipping design as developed by Card et al. (2008a) and also used by Pan (2010). As depicted in Figure 1.1, the RD research strategy exploits the cross-sectional variation in base-year minority shares across workplaces to test whether workplaces exhibit tipping patterns as the initial minority share in a workplace exceeds a certain critical threshold. The location of the candidate tipping points is assumed to be sector-specific and is identified by a "fixed-point" procedure that builds on the shape of Figure 1.2. Figure 1.2 plots the mean net percentage changes in white employment in the construction sector (NAICS Sector 23) from 1995 to 2000 against the minority share in 1995. The horizontal line depicts the unconditional mean net white employment growth. The vertical line is the estimated tipping point using the "fixed-point" procedure elaborated below. The figure shows clear evidence that, compared to an average single-establishment firm in the construction sector, white employment increases relative to minority employment to the left of the tipping point and decreases substantially to the right of the tipping point. In Appendix Figure 1.A, I show that similar patterns exist in a broad sample of sectors for

both of five-year intervals studied.

Unlike the work of Pan (2010), which is conducted using occupation-state cells and is agnostic about the level at which the tipping mechanism operates, this paper uses establishment-level data from the Census Bureau's LEHD infrastructure files. With these data, I can study employment segregation dynamics at the otherwise hard-to-observe workplace level, where I expect to produce more accurate estimates of the magnitude of segregation and the tipping effect. Moreover, the linked employer-employee data structure enables me to show the shifting racial and ethnic composition of employers at the workplace level. In this way, my study delineates potential mechanisms under which workers respond to changes in the minority composition of their employers.

To motivate my econometric analyses, I first use the Duncan and Duncan index of dissimilarity [3] to show that segregation exists in the sample of firms used in this study. Because the social interaction model relies on the explicit assumption that workers have perfect information about minority shares, I use only single-establishment firms in the analysis. [4] Thus, most of the firms in my sample are small- to medium-sized. For small firms, indices such as the Duncan and Duncan index and the Gini index, which are widely used to quantify segregation, tend to overestimate its true magnitude (Carrington and Troske, 1997, 1998). This issue was first elaborated by Blau (1977) in the gender segregation literature. The

[3]The Duncan and Duncan index of dissimilarity is a measure widely used to quantify the degree of segregation. It can be written as

$$D_{i-j}^{K} = \frac{\sum_{k=l}^{K} |X_i^k - X_j^k|}{2}$$

where i and j denote different demographic groups; X_i^k and X_j^k denote the percent distribution of group i and j in occupation/industry/firm k; therefore $\sum_i^k = 100$ and $\sum_j^k = 100$ hold. Basically, the value of the index indicates the percentage of workers in group i who must change occupations/industries/firms to achieve an occupational/industry-wide/firm distribution identical to that of the group j workers. The index takes values between zero and one. When it equals zero, it indicates that groups i and j have the identical occupational distributions, i.e. no segregation; when the index equals 100, it indicates that group i and j workers are never in the same occupation, i.e. complete segregation.

[4]One clarification is necessary before delving into details. In this paper, the definitions of firms and establishments follow Abowd et al. (2009) in which establishments are defined as the place where the employees actually perform their work, and firms are defined as the legal entities that employ workers. Thus, firms can either be single-establishment employers or multi-establishment employers. In the following sections, the terms workplace, establishment, and single-establishment firm share the same definition and are used interchangeably in this paper.

causes of this distortion are two-fold: first, an integer constraint exists in which each worker must be uniquely allocated to one unit; second, the random allocation of workers to units does typically generate some deviation from complete evenness for small firms (Blau, 1977). To address this problem, Blau (1977) develops a random worker-to-firm allocation model. Inspired by Blau's model, I first verify systematic workplace-level segregation by computing the actual and expected Duncan and Duncan indices. Then, I proceed to my tipping-point estimation.

Turning to the employment dynamics, I find that establishment-level segregation is widely evident at the end of both five-year time periods. Using the 2000 and 2005 establishment-level data from the LEHD infrastructure files, I find that, compared to whites, minorities are much more likely to work at firms with at least 50 percent minority employment. I further confirm the existence of systematic workplace segregation across all sectors in both years. The average estimated candidate tipping points, which are measured in base-year minority shares and are estimated using the fixed-point procedure, are 14.16 percent in 1995-2000 and 15.51 percent in 2000-2005. Heterogeneity in the locations of the candidate tipping points does exist by industry.

In summary, I find strong evidence confirming that tipping exists in both five-year intervals among the single-establishment firms in the sample and it is rather robust to adding flexible controls of establishment-level covariates. I also demonstrate that the observed tipping pattern is mostly driven by non-Hispanic whites leaving. The effect due to minorities entering is small or even trivial. Such findings suggest that tipping patterns are associated with shrinking firms. That raises the concern that rather than social interactions, it may just be that whites are leaving firms that are not performing well. To address this concern, I restrict the analysis to establishments with minimal employment changes over each five-year period. Results show that the same tipping patterns still emerge. Alternative explanations, such as nonlinear changes in establishment characteristics, also fail to explain the observed effects. The tipping patterns described above are primarily found in service-producing sec-

tors rather than in the goods-producing sectors. Finally, I find that, unlike the 1995-2000 period, during which tipping behavior was driven equally by blacks and Hispanics, Hispanics are the sole driving force in the later 2000-2005 period. Taken together, this paper provides some of the first evidence suggesting that the dynamics of establishment-level segregation are highly nonlinear and exhibit a tipping pattern that is largely consistent with the Schelling (1971) social interaction model.

The paper is organized as follows. Section 1.2 lays out the model and the identification strategy and research design. Section 1.3 elaborates on the firm-level data from the LEHD infrastructure files, the unit of analysis, and the sample for this paper. In Section 1.4, a model of the random allocation of workers to firm developed by Blau (1977) is used to baseline the extent of racial and ethnic segregation in the sample. Section 1.5 shows the main empirical results on tipping. Robustness checks are also presented. In particular, section 1.5.4 goes beyond estimating tipping patterns to study the dynamics of the shifting composition of firms. The question is whether the observed tipping pattern is driven by white flight or by minority entry. In section 1.5.7, I explore various definitions of "minority" and examine whether these distinct racial and ethnic minority groups drive the tipping pattern differentially. Finally, Section 1.6 summarizes and concludes.

1.2 Model and Identification Strategy

My goal is to investigate the underlying mechanism that leads to workplace segregation by race and ethnicity. In particular, I want to test whether workplaces exhibit tipping patterns as the initial minority share in a workplace exceeds a certain critical threshold. The main analysis assesses whether social interaction models, as originally outlined by Schelling (1971), can account for the empirical evidence on nonlinear patterns of workplace minority composition changes. A brief review of Schelling's model (originally applied to residential segregation) is presented in Appendix 1.7.B. Schelling's tipping model has two key features:

(1) for tipping to occur, heterogeneity in preferences over neighborhood minority composition must exist; and (2) because the tipping point and the actual tipping are characterized as an unstable equilibrium and a dynamic adjustment process, there must be some friction that ensures that individuals do not always immediately go to the long-run stable equilibrium. In Schelling's model, this friction arises because individuals are myopic decision-makers (Caetano and Maheshri, 2014). Following the standard setup, the theoretical model presented in section 1.2.1 adopts these two key features as well.

A central insight of Schelling's model is that at any given point, neighborhoods may be observed in the process of tipping, i.e., in disequilibrium, rather than a stable long-run equilibrium. However, most current empirical neighborhood-choice models assume that household choices are observed in equilibrium. Models that are always in equilibrium cannot be used to implement empirical versions of Schelling's tipping model (Caetano and Maheshri, 2014). Card et al. (2008a) circumvented this problem using an approach that identifies a tipping point as a bifurcation point or threshold around which the flows of both whites and minorities are quantitatively different (Caetano and Maheshri, 2014; Card et al., 2008a,b). In other words, unlike Schelling's model, in which the only stable equilibria are complete segregation and the neighborhood tipping points are characterized as disequilibria, the tipping points in Card et al. (2008a) represents the maximum minority share at which a neighborhood can maintain a stable integrated equilibrium (Card et al., 2008a,b) which permits empirical identification. This paper builds on Card et al. (2008a). I present in this section a model of firm tipping and an identification strategy to estimate the tipping phenomenon at the workplace level. A direct empirical implementation examines whether evidence of discontinuous changes in workplace minority composition at candidate tipping points exists.

1.2.1 A Model of Firm Tipping

I present a simple, static, partial equilibrium model in which whites' labor supply to single-establishment firms depends on the share of minority workers in that firm. I assume ho-

mogeneity in the job positions. [5] To focus attention on workers' labor supply decisions, I assume that labor demand is constant and that employers are non-discriminating. Based on these assumptions, in a partial equilibrium, workers from different groups will be paid equal wages in the same firm. [6]

Assume that there are two types of workers with distinct racial and ethnic characteristics: non-Hispanic whites (W) and racial/ethnic minorities (M). Workers observe the wage offers posted by all firms. Workers have perfect information about the minority shares in each firm, which are denoted as $R_j = \frac{N_j^M}{N_j^M + N_j^W}$, where j indexes the firm, and N_j^M and N_j^W are the total employment of minorities and whites in firm j. Workers are utility maximizing agents who differ in their tastes and preferences for the minority share at their employers.

Due to the assumption of perfect information on wage offers and minority shares in each firm, i.e., (ω_j, R_j), worker i of type $t \in \{W, M\}$ solves the following problem:

$$\max U_i^t(\omega_j, R_j)$$
$$s.t.\ j \in \{1 \cdots J\}$$

where $U(\cdot, \cdot)$ is continuous and twice differentiable. The following first-order and second-order conditions also hold:

$$\frac{\partial U}{\partial \omega} > 0 \quad \& \quad \frac{\partial^2 U}{\partial \omega^2} < 0, \quad \forall i,\ t$$
$$\frac{\partial U}{\partial R} < 0 \quad \& \quad \frac{\partial^2 U}{\partial R^2} > 0, \quad \forall i,\ t$$

Workers are myopic in the sense that they make decisions based on the wage offers and

[5] According to Appendix 1.7.A., NAICS "groups establishments into industries based on the activity in which they are primarily engaged. Establishments using similar raw material inputs, similar capital equipment, and similar labor are classified in the same industry...." (`www.bls.gov/bls/naics.htm`). Since the analysis is conducted in a sector-specific manner, this assumption is not unreasonable.

[6] The implicit assumption here is that workers from different racial and ethnic groups are perfect substitutes. Though assuming non-discriminating firms is a strong assumption, Becker (1971) developed a model of employee discrimination showing that employees' tastes of discrimination alone can lead to labor market segregation. Additionally, even if an employer has a taste of discrimination, Blau (1977) argues that there are institutional constraints internal to a firm that place limits on the employer's ability to differentiate among individual workers.

minority shares they observe without taking into account the simultaneous decisions made by other agents. Let n_j^t denote the number of workers of type t who supply their labor to firm j. Then, n_j^t can be written as:

$$\begin{aligned} n_j^t &= \sum_i \mathbf{1}(i : j = argmax\ U_i^t(\omega_j, R_j), j \in \{1 \cdots J\}) \\ &= n_j^t(\omega_j, R_j) \end{aligned}$$

In this model the labor supply of type W and type M workers to firm j depends on the firm's wage rate ω_j and its share of minority workers, R_j. Given the continuity and monotonicity of the utility function, the inverse labor supply functions exist and are unique. Let $\omega_j^W(n_j^W, R_j)$ and $\omega_j^M(n_j^M, R_j)$ be the inverse labor supply functions. Taking $\omega_j^W(n_j^W, R_j)$ as an example, "n_j^W" whites are willing to work in firm j with minority share R_j and wage ω_j^W. In a partial equilibrium with non-discriminating employers, fixed labor demand, and perfect substitutability, the following condition holds:

$$\omega_j^W(n_j^W, R_j) = \omega_j^M(n_j^M, R_j) \quad \forall\ j \tag{1.1}$$

To simplify the notation in what follows, the firm index j is dropped, but all the equations are derived at the firm level. Due to the construction of the inverse labor supply functions, $\dfrac{\partial \omega^W}{\partial n^W}$ and $\dfrac{\partial \omega^M}{\partial n^M}$ are weakly positive. The cross derivatives of the inverse labor supply function, $\dfrac{\partial \omega^W}{\partial R}$ and $\dfrac{\partial \omega^M}{\partial R}$, represent the social interaction effects. These interactions imply that whites require a premium to work with minorities in firms. This premium is assumed to be higher in firms with higher minority shares, i.e., $\dfrac{\partial \omega^W(n^W, R)}{\partial R} > 0.$ and $\dfrac{\partial^2 \omega^W(n^W, R)}{\partial^2 R} > 0.$

Under the assumption that labor demand is fixed and that employers are non-discriminating, I normalize the total number of workers in a firm to $\bar{L} = n^W + n^M = 1$. Given this normalization, in an integrated equilibrium with minority share $R \in (0, 1)$, we have the following

condition:

$$\omega^W(1-R,R) = \omega^M(R,R) \tag{1.2}$$

where $n^M = R$ and $n^W = 1-R$. The derivative of $\omega^W(1-R,R)$ with respect to the minority share is:

$$\frac{\partial \omega^W(1-R,R)}{\partial R} = -\frac{\partial \omega^W}{\partial n^W} + \frac{\partial \omega^W}{\partial R} \tag{1.3}$$

In equation (1.3), the first term is negative. With a positive social interaction effects, the white inverse labor supply function is unlikely to be monotonically increasing. If $\frac{\partial \omega^W}{\partial R}$ is small at $R = 0$ and becomes more positive as R increases, the white inverse labor supply function may initially be downward sloping. As the minority share rises, the positive social interaction effect will dominate, which leads to an upward-sloping inverse labor supply curve. [7] For illustrative purposes, $\omega^M(n^M, R)$ is assumed to be upward-sloping and linear. [8] The two inverse labor supply curves are depicted in Figure 1.3.

The firm depicted in Figure 1.3 has three equilibria: two integrated equilibria and one all-minority equilibrium. Point A is a locally stable integrated equilibrium. For instance, for any small perturbation to the right of point A, the marginal minority worker requires a higher wage than the marginal white worker, and the non-discriminating firm will therefore hire the marginal white worker, which will return the system to point A. Using similar reason, point B is not a stable equilibrium. Any positive shock at B will start the system trending toward the all-minority equilibrium C instead of back to B.

An increase in the supply of minority workers pushes the minority inverse labor supply function downward, as shown in Figure 1.4. Figure 1.4 illustrates a series of equilibria for this firm due to such a shift, assuming the white inverse labor supply function has the shape illustrated in Figure 1.3. At the low level of minority labor supply, $R = 0$ is a stable

[7] To ensure the existence of the critical point R^*, the social interaction function needs to be steeper than the function that characterizes the derivative of the own inverse labor supply curve, i.e. the following condition needs to be true: $\frac{\partial^2 \omega^W(n^W,R)}{\partial R^2} > \frac{\partial^2 \omega^W(n^W,R)}{\partial n^{W\,2}}$

[8] The derivative of the minority inverse labor supply function with respect to R is $\frac{\partial \omega^M}{\partial n^M} + \frac{\partial \omega^M}{\partial R}$; this could be downward if minorities have strong distaste towards all-white firms when R is low.

equilibrium (point A_0 in Figure 1.4). However, as the minority labor supply increases, i.e., ω^M shifts downward, wages begin to fall, and a few minority workers displace whites with the lowest willingness to supply. The firm will be in a stable integrated equilibrium (such as points A_1 and A_2 in Figure 1.4). Further increase in the supply of minority labor will cause the minority share to increase until ω^M is just tangent to ω^W. The minority share denoted as R^* is a "tipping point," representing the maximum minority share at which a firm can be in a stable integrated equilibrium. Once $R = R^*$, any further increase in minority labor supply will cause the integrated equilibrium to disappear and will lead to a fully segregated equilibrium (all-minority equilibrium, i.e., point D's in Figure 1.4). The location of the tipping point (R^*) depends on the strength of the social interaction effect.

Several points are worth emphasizing. First, notice that this model features a one-sided tipping pattern: firms with minority shares below the tipping point are potentially stable, but those that exceed the critical threshold rapidly converge to 100 percent minority composition. This contrasts with the classic Schelling model, which delivers a two-sided tipping outcome. [9] Second, my model delivers a tipping point even though white preferences for firm-level racial composition are continuous. In addition, wages evolve smoothly through the tipping point, even though employment shares change discontinuously. The smoothness of wages around the tipping point occurs because the upward-sloping minority inverse labor supply curve takes over smoothly from the white inverse labor supply curve at the discontinuity. Wages at the long-run $R = 1$ equilibrium can be higher or lower than at the tipping point depending on the shape of the minority inverse labor supply curves and their movements once tipping is underway.

[9] Using the census tract-level data from 1970 to 2000, Card et al. (2008b) find evidence that suggests tipping behavior is one-sided, and that minority composition in neighborhoods with initial minority shares below the tipping points stay relatively stable over time.

1.2.2 Empirical Implementation

Figure 1.4 assumes steady increases in relative minority labor supply (i.e. $\omega^M(R, R) - \omega^W(1 - R, R)$). On average, this is likely to be true because since the passage of the 1965 Immigration Act, the U.S. has experienced a new wave of immigration. These so-called "new immigrants" are mostly from less industrialized countries in South America and Asia (Xie and Gough, 2009). Due to firms' geographic dispersion and depending on the sectors which the firms belong to, there are likely to be firm-specific shifts in relative labor supply of whites and minorities. The model presented above explains how firm-level minority composition responds to these firm-specific shocks in relative minority labor supply. These insights can be broadly summarized with three scenarios:

(i) For a firm with an initial minority share R_{t-1} somewhat less than R^*, small shifts in relative minority labor supply will produce small changes in the location of the integrated equilibrium, and the firm will move smoothly toward the new integrated equilibrium, so long as the minority share remains below R^*. Formally, for the set of firms with initial minority share $R_{t-1} \in [0, R^* - s)$ where s represents the maximum relative minority labor supply shock between period $t-1$ and t, $E[\Delta R_t \mid R_{t-1}] = g(R_{t-1})$ for some continuous function $g(\cdot)$.

(ii) Firms with initial minority share above R^* have already begun tipping, the expected change in minority shares for such firms is going to be positive and large. Formally, for the set of firms with initial minority share $R_{t-1} > R^*$, $E[\Delta R_t \mid R_{t-1}] = h(R_{t-1}) > 0$.

(iii) The intermediate range, firms with initial minority share in $[R^* - s, R^*]$, will tip only if they experience sufficiently large shocks, but not otherwise.

Assuming s is very small, then the $E[\Delta R_t \mid R_{t-1}]$ can be written as follows:

$$E[\Delta R_t \mid R_{t-1}] = \mathbf{1}(R_{t-1} < R^*)g(R_{t-1}) + \mathbf{1}(R_{t-1} \geq R^*)h(R_{t-1}) \tag{1.4}$$

If $\lim_{\epsilon \to 0+} h(R^* + \epsilon) - g(R^* - \epsilon) > 0$, the right-hand side of equation (1.4) is discontinuous at R^* leading to a "jump." Given the nature of $g(\cdot)$ and $h(\cdot)$, such a jump is likely to be large. As a result, the empirical strategy is to test for a discontinuity in $E[\Delta R_t \mid R_{t-1}]$ at candidate values of R^*. Strictly speaking, a consequence of equation (1.4) is that for some firms, some time horizons, and some heterogeneity in the location of firm-specific tipping points the function $E[\Delta R_t \mid R_{t-1}]$ might not be strictly discontinuous at R^*. Instead, it will be very steep with a slope in the $[R^* - s, R^*]$ range. In this paper, such a pattern, if any, is also interpreted as evidence of tipping.

1.2.3 Empirical Strategy & the Identification of the Tipping Point

The empirical analysis uses data for single-establishment firms. I measure changes in their employment composition over a five-year interval. [10] Because the social interaction model relies on the explicit assumption that workers have perfect information about minority shares, I use only single-establishment firms. Let $W_{ijs,t}$, $M_{ijs,t}$ and $P_{ijs,t} = W_{ijs,t} + M_{ijs,t}$ denote the total numbers of whites, minorities, and total employment in firm i, industry j, state s and year t. The main dependent variable, which measures the establishment minority composition changes over a five-year interval, is the percentage change in net white employment,

$$Dw_{ijs,t} = \frac{(W_{ijs,t} - W_{ijs,t-5})}{P_{ijs,t-5}} - \frac{(M_{ijs,t} - M_{ijs,t-5})}{P_{ijs,t-5}} \tag{1.5}$$

In order to reveal the dynamics of the shifting composition of firms and document whether the observed tipping patterns are driven by white flight or minorities entering, I also examine

[10] In previous studies on residential and occupational segregation that employ a similar empirical strategy, 10-year changes calculated from the decennial census of population are usually used (Card et al., 2008a,b; Easterly, 2009; Pan, 2010). In this paper, instead of 10-year changes, I use five-year changes because: (1) workplace dynamics are more volatile compared to census tracts and occupations; and (2) data from the LEHD infrastructure files are collected more frequently than the population census data. This eliminates some of the data limitations faced by previous studies.

the analogous measures for whites and minorities, separately,

$$\frac{(W_{ijs,t} - W_{ijs,t-5})}{P_{ijs,t-5}} \; and \; \frac{(M_{ijs,t} - M_{ijs,t-5})}{P_{ijs,t-5}} \tag{1.6}$$

The key explanatory variable is the base-year minority employment share in a firm,

$$R_{ijs,t-5} = \frac{M_{ijs,t-5}}{P_{ijs,t-5}} \tag{1.7}$$

Equation (1.4) from section 1.2.2 implies that $E[Dw_{ijs,t} \mid R_{ijs,t-5}]$ is a smooth function of $R_{ijs,t-5}$ except, perhaps, at the tipping point R^*. In this paper, the tipping point, if any, is assumed to be industry specific because some industries may be more prone to minority inflows than others. For instance, in 1995, approximately 17 percent of total employment in the construction sector was black or Hispanic. This share increased to 21 percent in 2000 and to nearly 30 percent in 2005. In comparison, the percentage of blacks and Hispanics employed in finance, insurance, and real estate has remained between 16 to 18 percent since 1995. [11]

Denote $R^*_{j,t-5}$ as the potential tipping point for industry j in year $t-5$, let $\delta_{ijs,t-5} = R_{ijs,t-5} - R^*_{j,t-5}$ be the deviation in minority share of firm i from its industry specific tipping point. The basic empirical specification is:

$$Dw_{ijs,t} = \phi(\delta_{ijs,t-5}) + d\mathbf{1}[\delta_{ijs,t-5} > 0] + X_{ijs,t-5}\beta + \eta_j + \tau_s + \varepsilon_{ijs,t} \tag{1.8}$$

where $\phi(\cdot)$ is a smooth control function, modeled as a third-order polynomial; η_j is the fixed NAICS sector effect, τ_s measures the fixed state effect; $X_{ijs,t-5}$ is a vector of firm-level control variables. The contols including the share of workers who are at least 57 years old in the base period ($\%RET_{ijs,t-5}$). The age cutoff is set to be 57 years old because people

[11] Due to data limitations, only blacks and Hispanics are discussed here. Data are retrieved from the 1995, 2000, and 2005 Statistical Abstract data collected for the Statistical Compendia program (`http://www.census.gov/prod/www/statistical_abstract.html`). The data are collected from the section on Labor Force, Employment and Earnings.

of this age or older are at risk of retiring during the next five-year window. Age 62 is the earliest age that one can claim social security benefits for retirement. Many studies have confirmed the effects of social security benefits on the elderly labor supply. Firms might experience decreases in white employment simply because they have larger shares of workers who are close to retirement age. Firm-level controls also include the share of young workers ($\%YOUNG_{ijs,t-5}$). In this paper, young workers are defined as those who are 24 years old or younger in the base-period. Because young workers tend to change jobs more frequently, firms might experience large changes in minority composition simply because they have larger shares of younger workers. Finally, firm-level log average earnings ($\log \overline{e_{ijs,t-5}}$) are also controlled because workers may leave a firm simply because they find better pay elsewhere.

Unlike most research using the conventional RD design, in which the running variable [12] and the cutoff are clearly defined, a critical issue in estimating an empirical model like equation (1.8) is that the discontinuity point $R^*_{j,t-5}$ is unknown and must be estimated from the data. To elucidate the method used to obtain the candidate tipping point, assume, for the moment, that a tipping point do exists. The method used here, the so-called "fixed-point" procedure, is borrowed from Card et al. (2008a). This approach uses the shape of smoothed approximation to $E[Dw_{ijs,t} \mid j, R_{ijs,t-5}]$ for industries. Figure 1.2 reveals that firms that have not hit the industry-specific tipping point tend to experience greater-than-average growth in net non-Hispanic white employment; however, firms that have reached or exceeded the industry-specific tipping point tend to experience relative declines. Formally, this finding implies the following:

$$E[Dw_{ijs,t} \mid j, R_{ijs,t-5} = R^*_{j,t-5} - \xi] > E[Dw_{ijs,t} \mid j] > E[Dw_{ijs,t} \mid j, R_{ijs,t-5} = R^*_{j,t-5} + \xi] \quad (1.9)$$

for some $\xi > 0$. Thus, the industry-specific tipping point is the minority share at which the

[12] It is also known as the observed "assignment" variable that determines the treatment status in the RD literature (Lee and Lemieux, 2010).

white employment of a firm grows at the average rate for the industry. To identify this level, I first obtain a smooth approximation to $E[Dw_{ijs,t} \mid j, R_{ijs,t-5}] - E[Dw_{ijs,t} \mid j]$ and then solve for the root of this function, which is the industry-specific tipping point. [13] If the functional form is correct, this procedure will consistently estimate the location of the tipping points. A result in the structural break literature is that sampling error in the location of a change point (e.g., $R^*_{j,t-5}$) can be ignored when estimating the magnitude of the break (e.g., d) (Card et al., 2008a). I borrow this result and do not adjust the standard errors for the estimation of $R^*_{j,t-5}$.

1.2.4 Hypothesis Testing

Because equation (1.8) is estimated using the candidate tipping points located using the data, the estimates of d, $\widehat{d}$ will have a non-standard distribution under the null hypothesis that there is no discontinuity (Hansen, 2000). Card et al. (2008a) call this a specification-research bias problem. Conventional test statistics tend to reject the null hypothesis $d = 0$ too often. Hansen (2000) recommends comparing the estimates to a simulated distribution of $\widehat{d}$ under the null hypothesis that there is no discontinuity. Card et al. (2008a) propose a split-sample technique that uses a randomly selected sub-sample [14] to locate the tipping point and the remainder of the sample to estimate the magnitude of the tipping effect. The authors claim that because the two sub-samples are independent, estimates of $\widehat{d}$ from the second sub-sample will still have a standard distribution and will thus permit conventional hypothesis testing under the null hypothesis. In this paper, the split-sample technique is used to facilitate conventional hypothesis testing. I use a simple random 50 percent subset of my sample for the estimation of the tipping points. The remaining 50 percent is used for further econometric analysis.

[13] A detailed description on the "fixed-point" procedure can be found in Appendix 1.7.C. on Tipping Estimation.

[14] Two-thirds of the sample was used to locate the tipping points in Card et al. (2008a) because the "fixed-point" procedure is quite data-intensive.

1.3 Data & Sample

1.3.1 Firm-level Data and Unit of Analysis

The Longitudinal Employer-Household Dynamics (LEHD) infrastructure file system is a job-based longitudinal frame designed to represent the universe of individual-employer pairs covered by the state unemployment insurance system reporting requirement (with federal employees added in 2012). Information about employer characteristics is constructed using the Quarterly Census of Employment and Wages (QCEW). Demographic information about workers comes from two administrative data resources: the Person Characteristics File (PCF) and the Composite Person Record (CPR), which are sourced from administrative records. The longitudinally linked employer-employee structure of the LEHD data allows researchers to follow both workers and firms over time. Additionally, one can also identify workers who share a common employer in any given quarter. Firms in the LEHD data are defined by their state-level unemployment insurance account number. [15] Basic information about firms includes total payroll, firm size, firm age, geography, and industry. Information on individual demographic characteristics includes race, ethnicity, education, date of birth, sex, and place of birth. A more comprehensive overview and description of the LEHD infrastructure files can be found in Abowd et al. (2009).

To explore labor market segregation by race and ethnicity, there is a question of what the appropriate unit of analysis should be (Pan, 2010). Goldin (2002) finds that the "pollution" of occupational prestige by women may occur at the level of firms, occupations, industries, or within some sort of spatial boundaries such as cities, municipalities, or states. Due to the lack of availability and accessibility of firm-level datasets, most studies have focused on racial segregation at the level of occupations or industries. However, Hellerstein et al. (2008) found that racial and ethnic segregation at the three-digit industry level in the Decennial

[15] That is to say, for example, a Target in New York and a Target in New Jersey are considered different firms, but a Target in Ithaca, New York, and a Target in Binghamton, New York, are considered to be part of the same firm.

Employer-Employee Dataset (DEED) is usually one-third as large as the establishment-level segregation they document. They further assert that workplaces, i.e., establishment, should be the units of observation for studying labor market segregation since the essence of social interaction among workers is better captured at the workplace level.

Using the LEHD infrastructure files, this study can be conducted at the level of establishments or workplaces by considering only the single-establishment firms. Since the main dependent variable is the five-year change in non-Hispanic white employment as a fraction of the base-year total employment net of the minority fraction, this paper does not exploit the full longitudinal structure of the LEHD data but focuses on changes over two five-year windows: 1995 to 2000 and 2000 to 2005. These five-year windows were chosen to be consecutive and to cover a 10-year time span. The base year of the first five-year interval was chosen such that the sample covers a sufficient number of states. Since many states provide data to the LEHD program beginning in the mid-to-late 1990s, [16] I choose 1995-2000 to be the first five-year interval. To avoid any possible confounding impact due to the Great Recession, no further analysis is conducted for 2005-2010.

1.3.2 Sample

The sampling universe (frame), which is applied to both five-year intervals, is defined as follows: (1) firms must be private, non-farm (no NAICS sector 11) and non-public administration (no NAICS sector 92) firms; (2) firms must remain single-establishment in the base year and in the end year of a five-year interval; and (3) firms' establishment-level employment growth during a five-year interval must lie within 2.5 standard deviations of the state and NAICS sector averages for that time window. The purpose of restricting the sample in this manner is to avoid results driven by extreme values. The samples used for the 1995-2000 and 2000-2005 analyses are 50 percent simple random samples of establishments in each frame. The sampling procedure also selects the worker-level data for all individuals who are

[16] Detailed start dates for each state can be viewed at `http://download.vrdc.cornell.edu/qwipu/starting_dates.html`.

employed in their dominant job at the selected single-establishment firms in the base year and in the end year of a five-year interval.

Given the quarterly-based LEHD infrastructure files, there are many ways to construct the main dependent variable. In this paper, I used measures based the beginning-of-quarter employment in the second quarter [17] to construct the variables used in the empirical specification. The rationale is that the April 1 (the beginning of the second quarter) as the base for employment measures in a given year is closest to March 12, the reference date used by the Census Bureau for employment measures contained in its Business Register and in the Economic Censuses and Surveys (Abowd et al., 2009). A second rationale is that measures based on April 1 avoid discontinuities in the Unemployment Insurance wage records that occur at the change of calendar years.

In order to obtain the most economically meaningful results, the following sample restrictions are also applied. These restrictions are necessary because the earnings data in the LEHD infrastructure data are extracted from Unemployment Insurance covered earning records, in which any payment of at least one dollar made to an individual during the quarter will appear in the data. As a consequence, many one-time payments that do not necessarily agree with the general definition of employment between a firm and a worker appear as a "job" that lasts one quarter. Therefore, it is important to define a dominant job for a worker. Once the definition is formed, I consider a worker to be an employee only of her dominant-job firm. In this paper, I define a worker's dominant job in a year as the highest annual earning job for that year. Currently, individuals who have more than one dominant job (a small group who have identical earnings in two jobs over the year) or who indicate two or more races (a larger group) are excluded.

The final sample for 1995-2000 includes 200,000 unique single-establishment firms matched between 1995 and 2000 from 19 states, [18] 6,540,000 individuals in 1995, and 7,280,000 indi-

[17] Again, the definition of beginning-of-quarter employment follows Abowd et al. (2009)

[18] These 19 states include: CA, CO, FL, ID, IL, KS, LA, MD, MN, MO, MT, NC, NY, OR, PA, RI, TX, WA, and WI.

viduals in 2000. The final sample for 2000-2005 includes 341,000 unique single-establishment firms matched between 2000 and 2005 from 42 states, [19] 11,900,000 individuals in 2000, and 12,300,000 individuals in 2005. [20]

The individual characteristics file (ICF) in the LEHD infrastructure files contains all the necessary demographic variables used in this paper, including race, ethnicity, and date of birth. Approximately 3 percent of the individuals found in the unemployment insurance wage records do not link to the PCF [21](Abowd et al., 2009). To use effectively, the LEHD infrastructure files have undergone sophisticated multiple imputations using general Bayesian methods. [22] Ten independent missing data implicates are created to impute missing demographic variables for these individuals (Abowd et al., 2009). Each missing data implicate, combined with the observations with non-missing demographic information is referred to as an implicate file. To ensure the inference validity using the multiple imputation data, all the statistics and estimation are computed following Chapter 5 in Little and Rubin (2002). Each statistics or estimate is first computed 10 times using the 10 implicate files, individually. The final result is the mean estimand obtained by averaging across the results from the 10 implicate files. Standard errors are further corrected to account for missing data contribution to variance. [23]

[19]These 42 states include: AK, CA, CO, CT, DE, FL, GA, HI, IA, ID, IL, IN, KS, LA, MD, ME, MI, MN, MO, MT, NC, ND, NE, NJ, NM, NV, NY, OH, OK, OR, PA, RI, SC, SD, TN, TX, UT, VA, VT, WA, WI, and WV.

[20]These numbers are rounded to three significant digits for disclosure avoidance review purposes.

[21]As described in section 1.3.1, demographic information about workers comes from two administrative data resources: the Person Characteristics File (PCF) and the Composite Person Record (CPR).

[22]Refer to Little and Rubin (2002) for a detailed description of the general Bayesian methods for multiple imputation.

[23]Detailed computation formulas used in this paper are presented in Appendix 1.7.D.

1.4 Evidence on Systematic Firm-Level Segregation By Race and Ethnicity

1.4.1 Suggestive Evidence on Establishment-Level Segregation

Hellerstein et al. (2008) verify the existence of establishment-level segregation by race, ethnicity, and skills, using the Decennial Employer-Employee Dataset (DEED) in 1990 and 2000. This section shows that establishment-level segregation is still widespread at the end of each five-year window in the sample of firms used in this paper. Figure 1.5 is constructed to present the distributions of white and minority workers across single-establishment firms grouped by minority composition categories. This is done for all sectors pooled and for the goods-producing and service-producing NAICS supersectors, separately.

Overall, Figure 1.5 presents evidence suggesting that substantial establishment-level segregation is pervasive in 2000 and 2005. In particular, a comparison between the distributions of whites and minorities across various firm minority composition categories reveals a striking pattern: compared to non-Hispanic whites, minorities are much more likely to be employed in firms with higher minority shares. For instance, the top left figure, constructed for all sectors pooled in 2000, shows that approximately 3 percent of all minority workers work in firms where minorities account for less than 10 percent of the employment. Nonetheless, these firms account for close to 30 percent of all non-Hispanic white workers. In comparison, more than 30 percent of all minority workers work in firms where minorities account for 50 percent to 75 percent of the employment. This share remains high even when considering firms where more than 75 percent of the employment is minorities. On the other hand, these two groups of firms account for approximately 12 percent of all non-Hispanic white workers - approximately 10 percent in firms where minorities account for 50 percent to 75 percent of the employment and only about 2 percent in firms where minorities account for more than 75 percent of the employment. The all-sector pooled sample in 2005, which is depicted in the lower left figure in Figure 1.5, shows similar patterns.

Figure 1.5 also shows that the goods-producing and the service-producing supersectors in 2000 and 2005 exhibit patterns nearly identical to the one discussed above. A close comparison between the top and the bottom panels in Figure 1.5 indicates very minimal changes in the uneven distributions of whites and minorities between 2000 and 2005. To further illustrate that these trends and findings also exist in each NAICS sector, Appendix Figure 1.B is constructed using NAICS sector 23 (construction) and NAICS sector 62 (health care and social assistance), separately, as examples.

To examine whether various racial and ethnic minority groups exhibit different segregation patterns, I replicate Figure 1.5 for Asians, blacks, and Hispanics, separately. The results are presented in Appendix Figure 1.C. Although all three minority groups experience establishment-level segregation, blacks (represented by the red bars in Appendix Figure 1.C) seem to face the least. For instance, in 2000 and 2005, less than 50 percent of all black workers were employed in firms with 50 percent or higher minority shares in all sectors pooled. In both years, however, more than half of all Asian workers (represented by the blue bars in Appendix Figure 1.C) and Hispanic workers (represented by the green bars in Appendix Figure 1.C) were employed in these firms. Additionally, blacks have the highest proportion of workers at firms with less than 25 percent minorities in all sectors. By contrast, Hispanic workers have the lowest proportion. These findings also hold for the goods-producing and service-producing supersectors. Because Asians and Hispanics are the main immigrant groups in recent decades and, compared to blacks, have a much shorter history in the U.S., the results seem to suggest that these two minority groups might face more prejudice. [24]

1.4.2 Evidence on Systematic Firm-Level Segregation

A conventional way to document segregation is to compute the Duncan and Duncan index. As discussed in Section 2.1, when firm sizes are relatively small, the Duncan and Duncan index tends to distort the true magnitude of segregation (Blau, 1977; Carrington and Troske,

[24] Although it is entirely possible that these newer immigrant groups have not assimilated and therefore distribute more unevenly.

1997, 1998). The main cause of this distortion is that the conventional Duncan and Duncan index characterizes "no segregation" with an absolute zero value. However, research has shown that the segregation indices can be positive when workers are allocated randomly across units (Carrington and Troske, 1997, 1998). In an effort to address this concern, Blau (1977) developed a random worker-to-firm allocation model to adjust and allow complete randomness to be characterized by a non-zero benchmark Duncan and Duncan index.

Though Figure 1.5 provides suggestive evidence, it does not present any information on whether the observed pattern is systematically different from what would have been randomly observed by chance. To provide this information, I apply the random worker-to-firm allocation model developed by Blau (1977) to the same set of firms used in the previous section. [25] This model enables me to compute the distribution of firms that would have been observed by chance under the conditions of random worker-to-firm allocation, taking into account the minority composition of the labor pool for a state-NAICS sector. Then, this theoretical distribution of firms and the actual distribution can be used to compute the expected and the actual Duncan and Duncan index for each state-sector. Next, weighted averages of these two indices across all available states within each sector are computed. These sector-specific Duncan and Duncan indices for 2000 and 2005 are in Table 1.1. The expected Duncan and Duncan index defines the "evenness," and the difference between the expected and the actual Duncan and Duncan index measures the magnitude of systematic segregation. It is important to note that although I do not expect an absolute zero value in the Duncan and Duncan index to indicate evenness, as shown in Table 1.1, the expected Duncan and Duncan index is considerably less than the actual one.

As Table 1.1 demonstrates, a sizable proportion of minorities would have to reallocate among firms such that the actual distribution could be considered as indistinguishable from random worker-to-firm allocation. This statement holds for every sector. For instance, in 2000, close to 20 percent of minority workers in construction (NAICS 23) would have to

[25] The details of the random worker-to-firm allocation model are provided in Appendix 1.7.E.

reallocate among firms to approximate a situation of random allocations. In 2005, this index still remains higher than 18 percent. In both years, the sector that showed the most severe systematic segregation was health care and social assistance (NAICS 62). The Duncan and Duncan indices for this sector in both years are higher than 30 percent and have remained fairly constant between 2000 and 2005. Among all sectors listed, utilities (NAICS 22) has the smallest difference between the actual and expected Duncan and Duncan index for 2000 and 2005. Even then, for utilities to be considered a sector without systematic segregation, approximately 12 percent of minorities in 2000 and 14 percent in 2005 would have to be reallocated among firms. [26] Thus, Table 1.1 indicates that systematic segregation does exist at the establishment level in 2000 and 2005, although its extent appears to vary by industry. Nonetheless, the magnitude seems to vary minimally between 2000 and 2005. [27]

1.5 Do Firms Exhibit "Tipping-like" Patterns?

1.5.1 Descriptive Statistics

Table 1.2 presents descriptive statistics for the establishment-level data in all sectors pooled. The same descriptive statistics are also computed for establishments in the goods-producing supersector as well as in the service-producing supersector. The mean establishment-level minority shares in these two five-year intervals across sectors are very similar and are between 33% and 34%. In particular, Hispanics always comprise the largest minority group.

Overall, there is rapid employment growth in the period 1995-2000, which reflects the economic boom in the mid-to-late 1990s. As shown in Table 1.2, the goods-producing supersector and the service-producing supersector are equally affected by the economic boom. Although between 1995-2000, non-Hispanic white employment grows by more than 4 per-

[26] I have also applied the chi-square "goodness of fit" test developed by Blau (1977) to test whether the theoretical distribution of firms is systematically different from the actual distribution. Most state-sectors reject the hypothesis of random worker-to-firm allocation and thus confirm systematic segregation.

[27] To further visualize the pattern of systematic establishment level segregation, Table 1.1 is also converted into Appendix Figure 1.D.

cent, over 60 percent of the total employment growth is driven by growth in minority employment. This is true for all sectors pooled, the goods-producing supersector, and the service-producing supersector. Hispanic employment experiences the largest growth compared to the other racial and ethnic minority groups. In comparison, total employment growth between 2000 and 2005 is considerably slower. The goods-producing supersector even experienced contraction, which reflects the economic recession that occurred in early 2000 and the loss of manufacturing jobs in the U.S. In all sectors pooled, almost all employment growth can be attributed to minority employment growth. Specifically, between 2000-2005, total employment grows by 3.49 percentage points, and 3.10 percentage points are due to growth in minority employment. Interestingly, only non-Hispanic whites and blacks experience employment contraction in the goods-producing supersector, with the former being close to -3.8 percentage points. As in 1995-2000, Hispanics undergo the largest employment growth in 2000-2005 compared to the other minority groups.

Table 1.3 compares five subgroups of establishments defined by the fraction of minority shares in the base year, i.e., 1995 or 2000. Table 1.3 shows how the growth in non-Hispanic white employment is affected by the base-year minority share. Taking all sectors pooled in 1995 as an example, one can see clearly from Table 1.3 that in establishments that have minority shares from 0 to 5 percent, more than 70 percent of the growth in total employment is driven by the growth in white employment. Establishments that were 5 to 20 percent minority saw relatively slower growth in white employment. Nonetheless, growth in non-Hispanic white employment accounts for approximately one half of total employment growth. In contrast, establishments that were 20 to 50 percent minority experienced much slower growth in white employment, although the magnitudes of total employment growth are not dramatically different compared to establishments with lower minority shares. When base-year minority shares further increase, growth in white employment remains low.

The findings here suggest that once the establishment-level minority share reaches a certain level in the base year, non-Hispanic white employment growth over the five-year

window tends to dramatically slow down. Because there is no such indication on total employment growth, the summary statistics presented in Table 1.3 imply that once the base-year minority share reaches a threshold level, minority composition increases dramatically, i.e., the tipping phenomenon occurs. It can be seen that the described pattern and trends hold true for all sectors listed in Table 1.3 except for the goods-producing supersector in 2000-2005. The nonconformity of the goods supersector may be due to the loss of manufacturing jobs during the recession in the early 2000s. Additionally, these trends generally remain true for all sectors, individually. To illustrate this finding, Appendix Table 1.A reproduces Table 1.3 for the construction and health care and social assistance sectors separately.

1.5.2 Pooled Analysis of Changes in Net Non-Hispanic White Employment Growth

In order to implement the RD-tipping design and estimate the empirical specifications developed in section 1.2.3, I use the fixed-point procedure first to obtain the candidate tipping points. The estimated sector-specific tipping points for 1995-2000 and 2000-2005 are presented in Table 1.4. These candidate tipping points range from 5.26 to close to 40 percent in 1995 and 2.44 to 38.6 percent in 2005. The mean tipping point across 18 sectors is 14.16 percent in 1995 and 15.51 percent in 2000. The increase in the average tipping point from 1995 to 2000 suggests an increasing level of tolerance for working with minorities in the same firm, although the increase is quite small.

I now turn to specifications that pool the data in all sectors but estimated separately for the 1995-2000 and 2000-2005 periods. Figure 1.6 depicts the relationship between the base-year minority share in a single-establishment firm, deviated from the sector-specific candidate tipping point, and the percentage change in the net non-Hispanic white employment in the establishment, deviated from its sector-specific mean. The dots in the figure represent mean changes in one-percentage bins of $\delta_{ijs,t-5} = R_{ijs,t-5} - R^*_{j,t-5}$. The solid green line is a local linear regression fitted separately on each side of the candidate tipping point with an

Epanechnikov kernel and a bandwidth of 5. Finally, the solid blue line shows fitted values from a global third-order polynomial in $\delta_{ijs,t-5}$, allowing an intercept shift at $\delta_{ijs,t-5} = 0$. I limit attention to $\delta_{ijs,t-5} \in [-20, 20]$.

Figure 1.6 suggests establishment-level tipping. In particular, the Figure presents clear evidence of a discontinuous change in the minority composition when comparing establishments just below and just above the tipping point. Although visually telling, Figure 1.6 does not permit formal hypothesis tests and does not control for other establishment-level characteristics that might affect worker mobility, making it hard to determine whether the observed tipping behavior is due to differences in other covariates close to the candidate tipping points.

1.5.3 Formal Econometric Evidence on Establishment-level Tipping

Table 1.5 presents estimates of $\widehat{d}$ from equation (1.8) pooling all sectors pooled in 1995-2000 and 2000-2005. The regression analysis assesses the magnitude of tipping for establishments with an initial minority share just above the sector-specific candidate tipping points, compared to establishments with an initial minority share just below the tipping points. The main dependent variable is the change in net non-Hispanic white employment over a five-year window as a percentage of the establishment total employment in the base year (columns (1) and (2)). To reveal the dynamics of the shifting composition of firms and to document whether the observed tipping pattern is driven by white flight or by minority entry, I also examine analogous measures for whites and minorities, separately (columns (4) and (5) for non-Hispanic whites; columns (6) and (7) for minorities; the results are discussed in detail in section 1.5.4).

The estimation controls for a flexible control function in a form of third-order polynomial in $\delta_{ijs,t-5}$, establishment-level covariates as described in section 1.2.3, fixed state effects, and fixed sector effects. Standard errors are clustered on the state-sector level. All estimates are

computed and averaged across the 10 implicate files. The variance-covariance matrices of the estimates are corrected by taking into consideration of the variance contribution of the missing data and multiple imputation. [28] The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

The estimated coefficients for the models in columns (1) and (2) confirm that the change in net non-Hispanic employment as a percentage of the establishment total employment is discontinuous in the initial minority share around the candidate tipping points. When I estimate the model without any establishment controls (column (1)), the estimated, statistically significant, discontinuities are approximately -6 and -3 percentage points in 1995-2000, and 2000-2005, respectively. In 1995-2000, other things equal, the growth in net non-Hispanic white employment in establishments with an initial minority share just above the sector-specific candidate tipping points is 6 percentage points less than in establishments with initial minority shares just below the tipping points. In 2000-2005, the discontinuity is also statistically significant, although the magnitude decreases to -3 percentage points. When establishment controls are included (column (2)), the estimated discontinuities in both five-year intervals remain largely unchanged. [29]

Column (3) in Table 1.5 presents estimates where the dependent variable is the change in the establishment's minority share, i.e., $R_{ijs,t} - R_{ijs,t-5}$. The estimated tipping effect on this variable, which is the traditional focus of tipping models (Card et al., 2008a; Easterly, 2009;

[28] The computation formulas can be found in Appendix 1.7.D.

[29] I re-estimated equation (1.8) by adding another establishment-level covariate: share of foreign-born workers. The concern is that maybe it is nationality rather than race and ethnicity that is causing this observed threshold effect. This is a reasonable speculation as since the mid-1990s foreign-born workers are much more likely to be racial and ethnic minorities as compared to native-born workers. It turns out that even with this addition control, the results described here remain largely unchanged.

Additionally, because people tend to live close to where they work, another major concern is that what I am capturing here is actually residential tipping as studied in Card et al. (2008a) rather than workplace tipping effect. To examine this hypothesis, I re-estimated equation (1.8) by adding in another covariate: five-year change in county-level minority share of those who are at least 15 years old. This variable is calculated using the county population estimates produced by the Census Bureau's Population Estimates Programs. Once calculated, it is merged in using firms' county FIPS code. Basically, after controlling for five-year change in county-level minority shares, although the magnitudes of the coefficients become slightly smaller, the significance level remain unchanged. In summary, changes in neighborhood minority composition fall short in explaining the workplace tipping effect by race and ethnicity I have presented here.

Pan, 2010), is significant in both five-year intervals, although it is larger in 1995-2000. For instance, the estimate obtained for 1995-2000 implies that, compared to the establishments with initial minority shares just below the candidate tipping points, there is a significant increase in minority share of more than 2 percentage points in establishments with initial minority shares just above the tipping points.

One possible reason that the estimated tipping effect decreased between 1995-2000 and 2000-2005 might be the effect of changes in the state composition of my main estimation sample. The 1995-2000 sample includes single-establishment firms from only 19 states compared to 42 states covered in the 2000-2005 sample. To verify that my results are not due to the change in the number of states covered in the second five-year interval, I replicate columns (2), (3), (5), and (7) in Table 1.5 for all 19 states from the 1995-2000 sample in 2000-2005. Table 1.6 presents these results. Table 1.6 shows that the magnitude of the observed discontinuity is indeed smaller compared to 1995-2000, even when using the same 19 states in 1995-2000 and 2000-2005.

Another possible explanation for the estimated decrease might be the recession that occurred in 2001 and the associated drops in quits and total separations. [30] Research has shown that worker churning and job-to-job mobility during recent recessions have declined considerably (Kahn and McEntarfer, 2013). In Appendix Figure 1.E, I plot the seasonally adjusted time-series data on quits and total separations of private establishments from the Job Openings and Labor Turnover Survey (JOLTS) produced by the Bureau of Labor Statistics (BLS). The Appendix Figure 1.E confirms that significant decreases in quits and total separations did occur in the 2001 recession (marked as the first shaded area in Appendix Table 1.E). The levels of quits and total separations remained fairly low until late 2003 and early 2004. In a different paper, Kahn (2010) finds that the cohorts who graduate from college in a bad economy also tend to have slightly higher tenure.

Thus far, all the analyses have been conducted using the pooled sector samples. Het-

[30] For the accurate start and end date of this recession, refer to `http://www.nber.org/cycles.html`

erogeneity in the tipping effect almost surely exists across different sectors. To explore this issue, Table 1.7 presents the results of applying the RD models to the goods-producing and services-producing NAICS supersectors, separately. The specifications are otherwise identical to those in columns (2), (5), and (7) in Table 1.5. It is clear that the observed tipping phenomena seems to exist only in the services-producing supersector. Discontinuity in the goods-producing supersector does not seem to exist in either five-year window. To further confirm this finding, Figure 1.7 plots the change in the net non-Hispanic white employment in the establishment, deviated from the sector-specific mean in the services-producing supersector in 1995-2000 and 2000-2005. These figures are identical in structure to Figure 1.6. In comparison, the pattern in Figure 1.6 is almost indistinguishable from that in Figure 1.7 for both five-year intervals, reinforcing the finding that the observed discontinuity exists only in the service-producing NAICS supersector.

1.5.4 Whites Leaving or Minority Entering?

The evidence presented thus far is consistent with the social interaction model and the tipping argument. However, there are alternative mechanisms that could also lead to a tipping phenomenon, such as changes in production technology or learning dynamics (Pan, 2010). For instance, the production technology argument suggests that the increase in minority labor supply into the labor market might lead firms to switch to a minority-intensive production technology, which could result in a sharp increase in minority employment growth over some range of initial minority share. Alternatively, a learning-dynamics model implies that at low minority shares, little information about a particular job is available, hence minority employment growth is slow. As the minority share rises, information accumulates and learning accelerates, which could lead to a rapid increase in minority employment (Pan, 2010).

A common way to try to distinguish these models from the social interaction model is to consider whether establishment-level tipping is driven by white flight or by minorities

entering. Schelling's mechanism suggests that tipping should be driven primarily by a sharp decline in non-Hispanic white employment, although it is entirely possible that minority employment might increase substantially in response to white flight. Nonetheless, if we observe a sharp decline in non-Hispanic white employment that is not accompanied by a sharp increase in minority employment, this would suggest that tipping is driven mostly by the social interaction model.

To examine the shifting composition of firms and, in particular, to examine whether establishment-level tipping is driven by white flight or minority entry, columns (4)-(7) in Table 1.5 present models for the changes in white and minority employment as a percentage of base-year total establishment employment for 1995-2000 and 2000-2005. The specifications are otherwise identical to those in columns (1) and (2) in Table 1.5. Columns (4)-(7) show that in 1995-2000, there was a significant decline in white employment growth and an upward jump in minority inflows at the sector-specific tipping points. In other words, the observed discontinuity in minority composition during this five-year window is driven almost equally by whites leaving and minorities entering, although the magnitude of the former is slightly larger. In comparison, in 2000-2005 the observed tipping effect is driven solely by whites leaving, and the upward jump in minority employment at the candidate tipping points is negligible. These results indicate that although tipping is confirmed to be a mechanism leading to establishment-level segregation in the sample of firms used in this paper, I cannot rule out that multiple explanations might explain the observed tipping phenomenon. For instance, while both the production-technology and learning-dynamics models would have trouble explaining the negligible effect due to minority entry in 2000-2005, I cannot reject a possible role for these hypotheses in the earlier five-year window (1995-2000). Therefore, it is entirely possible, and even likely, that more than one underlying process is operating. The purpose of this paper is to document the tipping patterns and to demonstrate that at least some of these observed patterns are broadly consistent with predictions from a simple Schelling-type social interaction model.

1.5.5 Does Tipping Only Exist in Shrinking Firms?

The specifications in Table 1.5 show that tipping is associated with a discontinuous drop in non-Hispanic white employment growth but a smaller or negligible jump in minority employment growth. Such findings suggest that tipping patterns are associated with shrinking firms. That raises the concern that rather than social interactions, it may just be that whites are leaving firms that are not performing well. Similarly, in Table 1.3 I find that establishments with initial minority shares below 20 percent experienced faster employment growth over the next five years compared to those with higher initial shares. These observations call attention to an important element that is missing from the model: labor demand.

With fixed labor demand, as in the model in Section 1.2, any decline in white labor supply is mechanically offset by minority labor inflows. To approximate an environment of fixed labor demand, I identify a subset of establishments where total employment has changed by less than 10 percentage points over a five-year interval. The model specification is otherwise identical to Table 1.5. The results are presented in Table 1.8.

In the establishments with fixed labor demand, the estimated discontinuity in net non-Hispanic white employment growth at the tipping point is -4.24 percent in 1995-2000 and -1.87 percent in 2000-2005. This is somewhat smaller than the corresponding estimate from the full sample (Column (2) in Table 1.5) but is still large and statistically significant. In these establishments with fixed labor demand, total employment growth shows no discontinuity at the tipping point, while the estimated discontinuity in white employment growth is approximately equal and opposite to the jump in minority employment growth. This observation is true for both 1995-2000 and 2000-2005. Column (2) of Table 1.8 presents estimates where the dependent variable is the change in establishment's minority share. The estimated tipping effect on this variable is apparent, although it is rather small in 2000-2005. Thus, Table 1.8 demonstrates that mobility patterns in these establishments with fixed labor demand closely match the predictions from the model with fixed labor demand.

1.5.6 Omitted Variables and Effect on Establishment Covariates

An additional concern with the RD model used in the previous sections is that the discontinuous relationship between net white mobility flows and the initial minority share might be due to omitted establishment characteristics that happen to be discontinuously related to the minority share. Although the main specifications (columns (2), (5), and (7) in Table 1.5) include a vector of establishment controls, these linear controls might not be flexible enough to absorb the nonlinear effects. To assess this possibility and to test whether the results presented in Table 1.5 are sensitive to flexible controls for the pre-period establishment characteristics, Table 1.9 presents a series of extended specifications that add a third-order polynomial in these establishment-level covariates. Table 1.9 shows that the estimates of $\widehat{d}$ are rather robust to such inclusions, suggesting that omitted variables of this kind are unlikely to account for the observed discontinuities.

The empirical analysis thus far has focused on changes in minority composition due to non-Hispanic white or minority employment growth. In other words, the analysis has primarily looked at changes in quantities. Nonetheless, apart from quantities, there are other outcomes worth examining. These include whether earnings, the share of retiring workers, or the share of young workers are affected by tipping. This part of the analysis therefore looks at how these establishment-level characteristics behave around the sector-specific candidate tipping points.

Table 1.10 reports results from regressions in which the dependent variable is replaced by changes over a five-year window in log average earnings for all workers, in the share of retiring workers, and in the share of young workers. In each regression, a flexible third-order polynomial in $\delta_{ijs,t-5}$, fixed state effects, and fixed sector effects are controlled. Similarly, standard errors are clustered on the state-sector. According to Table 1.10, there is little evidence of significant changes in the establishment-level covariates around the candidate tipping points. Thus, from Table 1.10 I conclude that the observed discontinuity is not

driven by abrupt changes in establishment-level characteristics around the tipping points. [31]

1.5.7 Minority Definition

Thus far, I have defined minorities as nonwhites and white Hispanics. Nonetheless, it is entirely possible that whites might react to inflows of different minority workers differently. Specifically, evidence presented in Appendix Figure 1.C suggests that different racial and ethnic minority groups seem to face different degrees of segregation and Table 1.2 shows clear heterogeneity in employment trends across Asians, blacks, and Hispanics. In this section, I present a series of models in which I vary the definition of minority to explore this issue.

Tables 1.11 and 1.12 present estimates that explore alternatives that count only blacks or only Hispanics as minorities for 1995-2000 and 2000-2005, separately. I also present a composite model that includes indicators for being beyond the tipping point for all three minority definitions. As in earlier tables, the dependent variable in each specification is the change in net non-Hispanic white employment, the change in white employment, or the change in minority (all nonwhites and white Hispanics) employment, as a percentage of total establishment employment. Candidate tipping points are estimated separately for each definition of minorities, using the fixed-point procedure discussed in section 1.2.3 and Appendix 1.7.C. Each model also includes a third-order polynomial in the deviation of the establishment's minority share from the candidate tipping point. The composite model includes all three third-order polynomials. The establishment controls are identical to those in Table 1.5. Fixed State and sector effects are included. Standard errors are clustered on the state-sector.

The estimates in columns (1)-(4) in Table 1.11 suggest that in 1995-2000, tipping behavior was driven slightly more by the black shares than by the presence of other minority

[31] Unfortunately, this is not true when I use the five-year change in the share of foreign-born workers as the dependent variable to conduct this robustness check. In other words, I cannot completely rule out the possibility that the workplace tipping effect may be driven by discontinuity in the share of foreign-born workers at the establishment-level.

groups, although the effect of the Hispanic shares is quite strong as well. When I decompose the observed discontinuity in net non-Hispanic white employment change into changes in white employment and minority employment and examine them separately, Table 1.11 (columns (5)-(8)) shows that changes in non-Hispanic white employment alone exhibit even stronger tipping beyond the black-share fixed point and the Hispanic-share fixed point. Interestingly, the results in Table 1.11 (columns (9)-(12)) seem to show that even minority workers leave establishments once its black share or Hispanic share reaches the candidate tipping points, with both measured the same way. However, the latter discontinuity is much smaller compared to the former.

In comparison, the results presented in Table 1.12 imply that in 2000-2005, almost all observed tipping behavior is driven solely by the Hispanic share. When I look at the white employment change and the minority employment change individually, the Hispanic shares seem to be the only driving force again. In particular, non-Hispanic white employment in establishments with initial Hispanic share just above the sector-specific candidate tipping points, measured in Hispanic share in 2000, experience a 7 percentage point decrease (column (7) in Table 1.12) compared to establishments with shares just below the tipping points. The discontinuity observed in minority employment change is much smaller, although it is statistically significant and of approximately -2.5 percentage points in magnitude (column (11) in Table 1.12). In all, estimates presented in Table 1.11 and 1.12 suggest that as Hispanics become the largest minority group in the U.S., they might face stronger distaste from non-Hispanic whites, and such distaste might even exist among other ethnic minority groups.

1.6 Conclusion

In summary, using the establishment-level data from the LEHD infrastructure files and the random worker-to-firm allocation model developed by Blau (1977), this study first confirms

that systematic racial and ethnic workplace segregation exists in 2000 and 2005. Then, the paper makes use of a Regression Discontinuity design developed by Card et al. (2008a) and demonstrates the importance of tipping. My approach uses the cross-sectional variation in base-year minority shares across establishments to test whether establishments exhibit tipping-like behavior in response to firm-specific shocks in minority labor supply that occur over two five-year intervals: 1995-2000 and 2000-2005.

The average NAICS sector-specific candidate tipping point, estimated using the fixed-point procedure, is 14.16 percent in 1995 and 15.51 percent in 2000. The increase in the average tipping point from 1995 to 2000 suggests an increasing tolerance level for working with minorities in the same firm, although this increase is quite small.

Overall, I find clear evidence that tipping is a feature of the dynamic process of establishment-level segregation in the sample of firms used in this paper. The estimated, statistically significant, discontinuities are close to -6, and -3 percentage points in 1995-2000, and 2000-2005, respectively. One possible reason for the decrease in the observed tipping effect between 1995-2000 and 2000-2005 is the recession that occurred in 2001. To examine the shifting composition of firms, and in particular, to explain whether establishment-level tipping is driven by white flight or minorities entering, I find that tipping in 1995-2000 is driven by whites leaving and minorities entering together. In comparison, in 2000-2005, the observed tipping effect is solely driven by white leaving, the upward jump in minority employment at the candidate tipping points is quite negligible. By using a subset of establishments that have undergone minimal employment growth over a five-year window to approximate an environment with fixed labor demand, I demonstrate that mobility patterns in these establishments closely match the predictions from the social interaction model with fixed labor demand presented in this paper. Taken together, the analysis in this paper provides some of the first evidence suggesting that the dynamics of establishment-level segregation are noticeably nonlinear and exhibit a tipping pattern. This observation is largely consistent with the Schelling (1971) social interaction model, although at this point, I cannot completely rule

out alternative explanations for the observed discontinuity. It is possible, and even likely, that more than one underlying process is operating here. Future work should assess how social interactions interact with other underlying mechanisms, which will provide a far richer picture of the dynamics of workplace segregation.

As part of robustness checks, I present evidence confirming that the tipping effects for both five-year windows are robust to adding flexible controls of establishment-level covariates. I also demonstrate that the observed tipping patterns are not driven by nonlinear changes in establishment characteristics. Finally, I present composite model estimates in which I explore alternative definitions of minority. In particular, I find that in 1995-2000, tipping behavior seems to have been driven slightly more by the black shares than by the presence of other minority groups, although the effect of the Hispanic shares is quite strong as well. In comparison, in 2000-2005, the observed tipping behavior seems to be driven solely by Hispanic shares. This change seems to suggest that as Hispanics become the largest minority group in the U.S., they might face stronger distaste from non-Hispanic whites. As the minority composition in the U.S. changes, this finding has implications for understanding the persistence of current labor market segregation.

1.7 Appendix

1.7.A. NAICS Sectors and NAICS Supersectors

NAICS stands for the North American Industry Classification System. Developed using a production-oriented conceptual framework, NAICS "groups establishments into industries based on the activity in which they are primarily engaged. Establishments using similar raw material inputs, similar capital equipment, and similar labor are classified in the same industry. In other words, establishments that do similar things in similar ways are classified together" (`www.bls.gov/bls/naics.htm`). Revisions implemented for every Economic Census (years ending in 2 and 7). In this paper, the 2007 NAICS classification is utilized. Overall, there are 20 NAICS sectors (`www.census.gov/cgi-bin/sssd/naics/naicsrch?chart=2007`).

For purposes of aggregate analysis, the U.S. Economic Classification Policy Committee aggregated NAICS sectors into "Supersectors." The goods-producing NAICS supersector includes natural resources and mining (NAICS 1133, i.e., logging; NAICS 21, i.e., mining), construction (NAICS 23) and manufacturing (NAICS 31-33) (`www.bls.gov/ces/cessuper.htm`). Because the sample in this paper does not include NAICS sector 11 (agriculture, forestry, fishing, and hunting), the goods-producing NAICS supersector only includes NAICS sectors 21, 23 and 31-33. The service-producing NAICS supersector includes trade, transportation, and utilities (NAICS 42, i.e., wholesale trade; NAICS 44-45, i.e., retail trade; NAICS 48-49, i.e., transportation and warehousing; NAICS 22, i.e., utilities), information (NAICS 51), financial activities (NAICS 52, i.e., finance and insurance; NAICS 53, i.e., real estate and rental and leasing), professional and business services (NAICS 54, i.e., professional, scientific, and technical services; NAICS 55, i.e., management of companies and enterprises; NAICS 56, i.e., administrative and waste services), education and health services (NAICS 61, i.e., educational services; NAICS 62, i.e., health care and social assistance), leisure and hospitality (NAICS 71, i.e., arts, entertainment, and recreation; NAICS

72, i.e., accommodations and food services), other services (NAICS 81), and government (`www.bls.gov/ces/cessuper.htm`). Because the sample in this paper does not include any governmental establishments, the service-producing NAICS supersector in this paper does not include government.

1.7.B. Schelling's Bounded-neighborhood Model

Schelling's Bounded-neighborhood model and its extension into the tipping model use the preference interaction perspective to analyze (residential) segregation by race (Schelling, 1971). Preference interaction occurs when an agent's preference ordering on the alternatives within her choice set depends on actions chosen by other agents (Manski, 2000).

In this model, there is a well-defined "neighborhood" with clear boundaries. People are either in or out of this common neighborhood. Everybody in this neighborhood is concerned with the minority share. This concern is characterized by a upper limit or tolerance for the minority share. An individual will reside in the neighborhood only if the minority share in the neighborhood has not reached his own limit. If an individual's limit is exceeded, he will leave and choose somewhere else that meets his tolerance level. This model assumes heterogeneity in individual preferences over the neighborhood-level minority share, ranging from complete integrationist to complete segregationist. Agents are assumed to have perfect information about the minority share within the neighborhood when they decide whether to leave or to enter a neighborhood. However, agents are myopic about other agents' intentions and their future moves. Zero mobility costs are also assumed. There are no neighborhood capacity constraints and adding-up constraints in the neighborhood to enforce that the population-weighted average of neighborhoods' minority shares be equal to the system-wide share of minorities in the population (Easterly, 2009; Schelling, 1971; Zhang, 2011). Therefore, Schelling's model cannot be viewed as a general equilibrium model.

Given this model setup, Schelling (1971) shows how only a modest preference of whites to live next to other whites can lead to nearly complete residential segregation. In this model, even a relatively small fraction of minorities could cause the neighborhood to tip from completely white to completely minority. The fraction at which this happens is called the "tipping point." The tipping point in Schelling's model represents an unstable equilibrium, since even a slight perturbation in the level of minority shares around the point can lead to complete segregation (Caetano and Maheshri, 2014). As a result, Schelling's model has the

feature that the only stable equilibria are fully segregated equilibria. A neighborhood with a mixed minority composition is inherently unstable. The triggered dynamic process can lead to either 0 percent or 100 percent minority share, i.e., two-sided tipping (Card et al., 2008b). A more detailed description of Schelling's Tipping model can be found in Schelling (1971).

1.7.C. Tipping Estimation

I use the fixed-point procedure discussed in section 1.2.3 to identify NAICS sector-specific tipping points in the 50 percent simple random subsample of establishments. I identify the roots of

$$E[Dw_{ijs,t} \mid j, R_{ijs,t-5}] - E[Dw_{ijs,t} \mid j] \tag{1.10}$$

as the estimated tipping point. I fit $Dw_{ijs,t} - E[Dw_{ijs,t} \mid j]$ to a third-order polynomial in $R_{ijs,t-5}$. Following Card et al. (2008a), I use only firms with minority shares below 60 percent. This polynomial is fitted separately for each NAICS sector. For each NAICS sector, I identify a root of this polynomial, taking into consideration the range of the minority shares in the remainder 50 percent subsample used for model estimation. In particular, I first exclude those roots above 50 percent minority share. The reason for restricting observations and the identified roots to this range is that this paper focuses on how establishments with lower shares of minorities in the base year respond to minority entry. Second, for each NAICS sector, I select roots that are strictly greater than the minimum value of base-year minority shares in establishments reserved for estimation. Finally, when there are multiple roots, the one that yields the most negative slope of the polynomial function is selected. The estimated sector-specific tipping point is presented in Table 1.4.

1.7.D. Computation Formulas for Multiple Imputation Statistics

This section follows Chapter 5 in Little and Rubin (2002). Let Y denote the data, which can be further partitioned into the observed and unobserved parts, if needed.

$$Y = (Y_{obs}, Y_{mis})$$

Let $Q(Y)$ denote the statistics of interest to be estimated. Let

$$Q_m(Y^m) = estimand\ from\ the\ m^{th}\ implicate$$

Let M denote the total number of implicates. Then, the average estimand over all implicates, $\overline{Q}$ can be written as

$$\overline{Q} = \frac{\sum_{m=1}^{M} Q_m(Y^m)}{M}$$

Let

$$V_m(Y^m) = covariance\ matrix\ of\ Q_m(Y^m)\ from\ the\ m^{th}\ implicate$$

Then, the average within-implicate covariance matrix, $\overline{V}$, can be written as

$$\overline{V} = \frac{\sum_{m=1}^{M} V_m(Y^m)}{M}$$

Let B denote the between-implicate variation of $Q_m(Y^m)$; then, B can be written as

$$B = \frac{[\sum_{m=1}^{M} (Q_m(Y^m) - \overline{Q})(Q_m(Y^m) - \overline{Q})^T]}{M}$$

The corrected covariance matrix, T, of $Q(Y)$, which accounts for the missing data contribution to variance, is defined as

$$T = \overline{V} + (1 + \frac{1}{M})B$$

The Rubin missingness ratio is defined as

$$Missingness\ Ratio = (1 + \frac{1}{M}) * \frac{b_{ii}}{t_{ii}}$$

where b_{ii} and t_{ii} are the diagonal elements of B and T. The Rubin missingness ratio essentially measures the proportion of the total variance that is due to between implicate variance. $\overline{Q}$, $\sqrt{t_{ii}}$, and the Rubin missingness ratio are the final results presented in all tables.

Within-implicate variance, i.e., V_m, for each estimate in Tables 1.1, 1.2, 1.3, Appendix Table 1.A, Figures 1.5, Appendix Figures 1.B, and 1.C is computed using the bootstrap method. The bootstrap samples for each implicate file m are generated by a simple random sampling with replacement, holding the sample size of that implicate file constant. The number of repetition is set to equal 1000. To compute the within-implicate variance for implicate file m, I first compute an estimand of interest, Q, for each bootstrap sample. Upon completion, the within-implicate variance of the estimand Q for implicate file m can be computed.

1.7.E. Random Worker to Firm Allocation Model

This paper adopts the random worker-to-firm allocation model developed by Blau (1977) and computes the expected and actual Duncan and Duncan indices for each NAICS sector presented in Table 1.1.

For each state and NAICS sector, let

p = the proportion of the individuals with the requisite industry-specific skills that is minority;

$q = 1 - p$ = the proportion of the labor pool that is non-Hispanic white;

x_i = the number of minorities employed in firm i in the given state and NAICS sector;

n_i = the total number of employees in firm i in the given state and NAICS sector;

$p_i = 100 * \frac{x_i}{n_i}$ = the share that minorities account for all workers in firm i in the given state and NAICS sector.

Under the random worker-to-firm allocation, x_i can be viewed as the outcome of n_i trials of an experiment in which each trial consists of selecting an individual at random from the labor pool, where the likelihood of getting a minority is p, and the likelihood of getting a non-Hispanic white is $q = 1-p$. Therefore, x_i can be characterized by a binomial probability distribution as:

$$f_i(x = x_i) = \binom{n_i}{x_i} p^{x_i} q^{n_i - x_i}$$

Then, firms are grouped according to size. Each size category contains firms with the same values of n_i. The possible outcomes, x_i, are grouped into ten categories according to the value of p_i: $0 \leq p_i < 10$, $10 \leq p_i < 20$, $20 \leq p_i < 30$, $30 \leq p_i < 40$, $40 \leq p_i < 50$, $50 \leq p_i < 60$, $60 \leq p_i < 70$, $70 \leq p_i < 80$, $80 \leq p_i < 90$, $90 \leq p_i \leq 100$.

Further, let:

n_j = the number of firms in the jth size category;

p_{jk} = the probability that a firm selected at random from the jth size class has a value of p_i that falls in the kth minority composition category;

e_{jk} = the expected number of firms in the j size class and kth minority composition category;

E_k = the total expected number of firms in the kth minority composition category;

P_k = the probability of obtaining a firm in the kth minority composition category, given the size distribution of firms.

Then, given the binomial probability distribution described earlier, p_{jk} can be written as:

$$p_{jk} = f(x_a \leq x \leq x_b) = \sum_{a}^{b} f_i(x = x_i)$$

Therefore, to find the theoretical distribution of firms with N firms in the state-two digit NAICS sector cell, e_{jk}, E_k, and P_k can be written as:

$$e_{jk} = p_{jk} \cdot n_j$$

$$E_k = \sum_{j} e_{jk}$$

$$P_k = E_k / N$$

The distribution of non-Hispanic white and minority workers among establishments that would prevail under the condition of random worker-to-firm allocation can be derived directly from the theoretical distribution of firms. Again, for each state and NAICS sector, let:

n_{ij} = the number of workers in firms included in the jth size class;

$\overline{p}_{ik}$ = the simple average of the p_i included in the kth minority composition category divided by 100;

m_{jk} and w_{jk} = the expected number of minorities and whites, respectively, employed in firms that fall into the jth size class and kth minority composition category;

M_k and W_k = the total expected number of minorities and whites, respectively, employed in firms included in the kth minority composition group.

Therefore, m_{jk} and w_{jk} can be approximated by

$$m_{jk} = e_{jk} \cdot n_{ij} \cdot \overline{p}_{ik}$$

$$w_{jk} = (e_{jk} \cdot n_{ij}) - (m_{jk})$$

And M_k and W_k can by calculated by the following:

$$M_k = \sum_j m_{jk}$$

$$W_k = \sum_j w_{jk}$$

Then, the state-sector-specific expected and actual Duncan and Duncan indices are calculated using the following formula:

Within each state and NAICS sector cell

Let p_i = the percentage that minority workers comprise the labor force in firm i. Then, firms are grouped into ten categories according to the value of p_i: $0 \leq p_i < 10$, $10 \leq p_i < 20$, $20 \leq p_i < 30$, $30 \leq p_i < 40$, $40 \leq p_i < 50$, $50 \leq p_i < 60$, $60 \leq p_i < 70$, $70 \leq p_i < 80$, $80 \leq p_i < 90$, $90 \leq p_i \leq 100$.$Let m_k$ and w_k equal the percentages of all minority workers and all non-Hispanic white workers who are employed in firms included in the kth minority

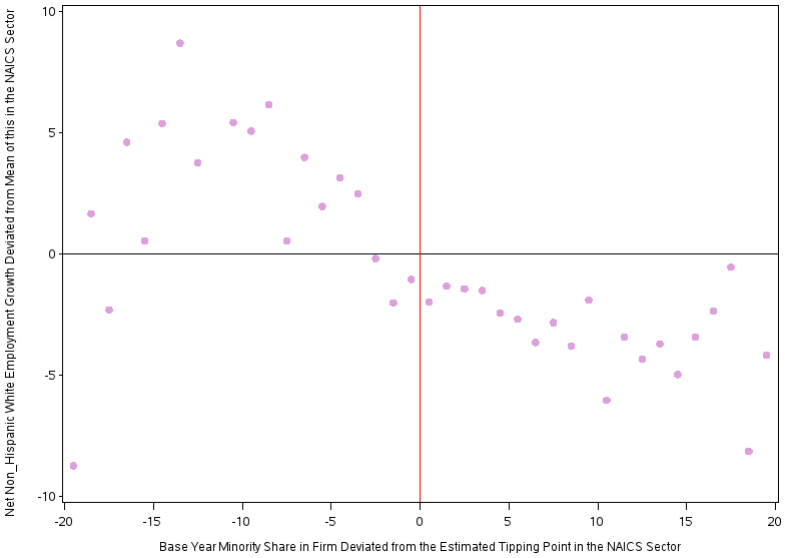
composition category. The Duncan and Duncan index of segregation for a given state and sector cell is defined as:

$$D = \frac{\sum_{k=1}^{10} |m_k - w_k|}{2}$$

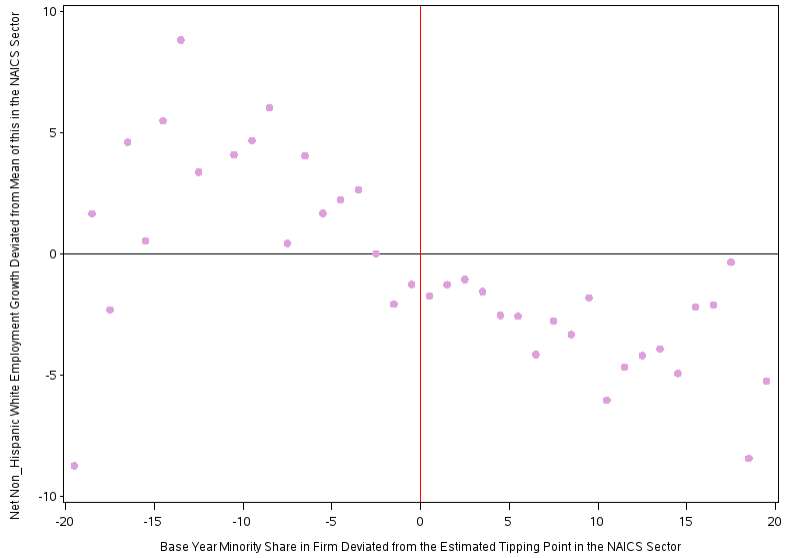
The actual Duncan and Duncan index of segregation is computed using the employment distribution of whites and minorities observed in the sample. The expected Duncan and Duncan index is computed using the theoretical distribution derived. Once the state and sector-specific indices are calculated, the NAICS sector-specific actual and expected indices are simply the weighted averages among all the available states. The weight used is the total number of firms in a given state-sector cell.

Figure 1.1. Change in a Pooled Sample of Firm-level Minority Composition, by Relationship to Candidate Tipping Points 1995-2000

All NAICS Sectors Pooled

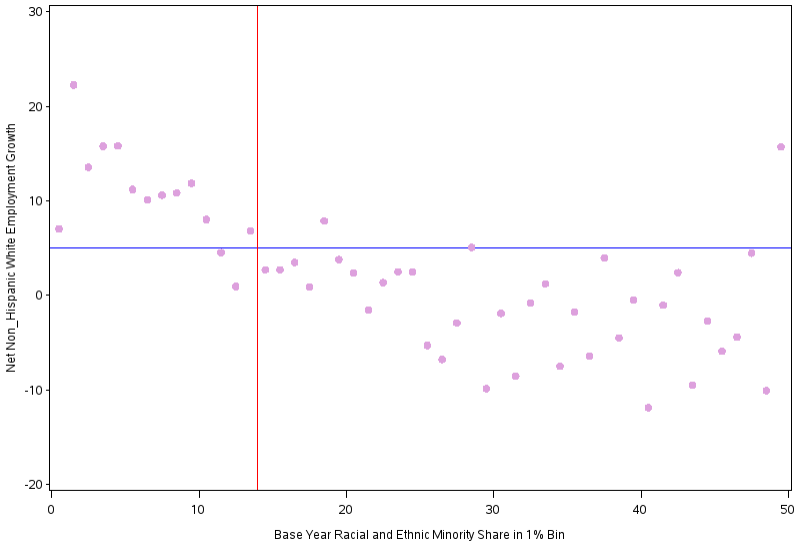


Service-producing NAICS Supersector



Notes: the X axis is minority share in establishment minus the estimated tipping point in a NAICS sector. The tipping point is estimated using the fixed-point procedure described in subsection 1.2.3. The Y axis is the percentage change in net white employment between 1995 and 2000, expressed as a percentage of the total establishment-level employment in 1995 and deviated from the mean in the NAICS sector. Dots depict averages in 1-percentage-point bins of the 1995 minority share. All series use only the 50% of establishments not used to identify the tipping points.

Figure 1.2. Firm-level Minority Composition Change in NAICS Sector 23 - Construction, 1995-2000



Notes: Dots show the mean of the change in the net establishment-level white employment between 1995 and 2000 as a percentage of the total employment in 1995, grouping establishments into cells of width 1% by the 1995 minority share. The horizontal line depicts the unconditional mean. The vertical line depicts the estimated tipping point using the fixed-point procedure described in subsection 1.2.3 and a 50 percent sample of single-establishment firms in NAICS sector 23.

Figure 1.3. Three Equilibria, With Social Interaction Effects

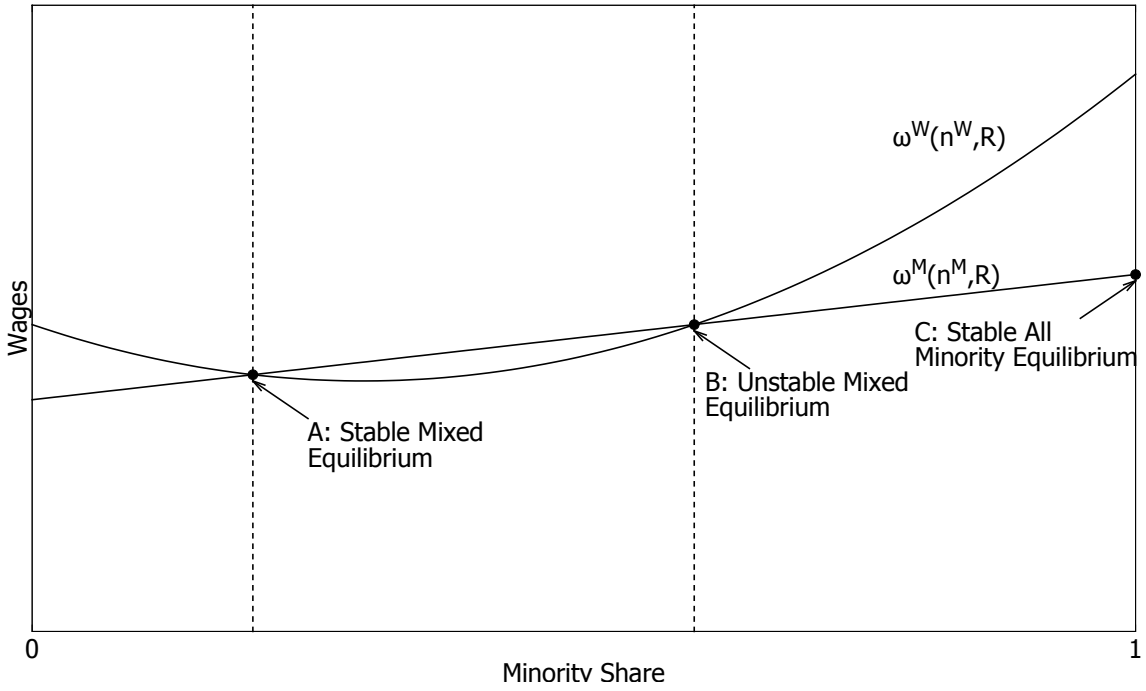


Figure 1.4. Rising Minority Labor Supply Leads to a Tipping Point

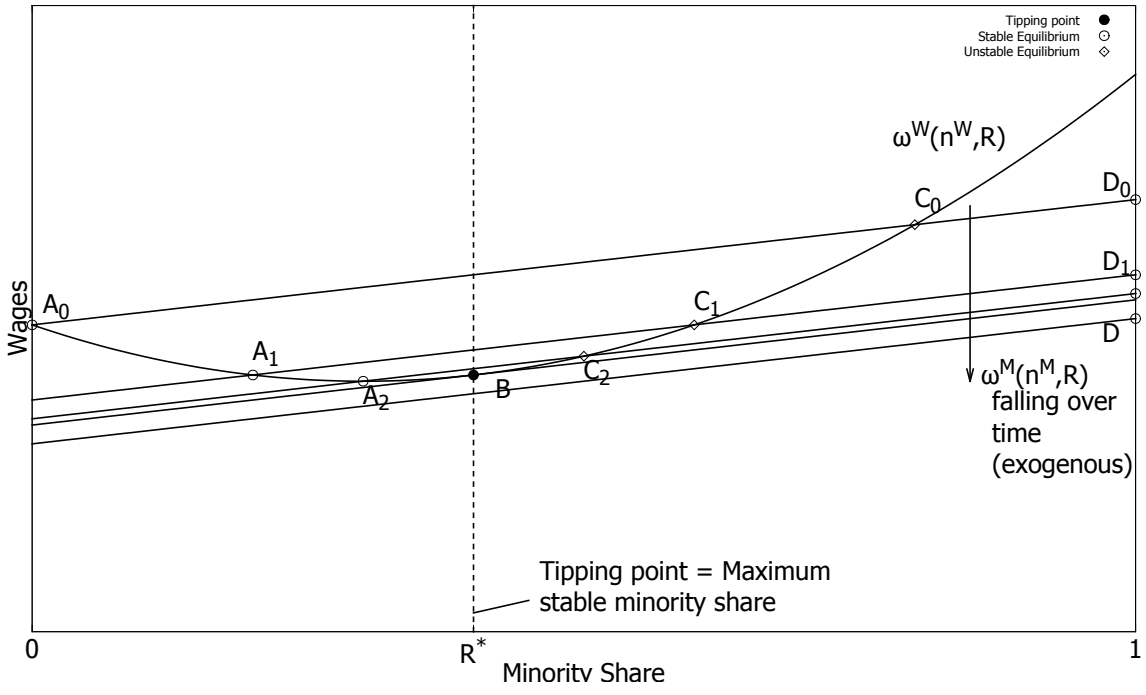
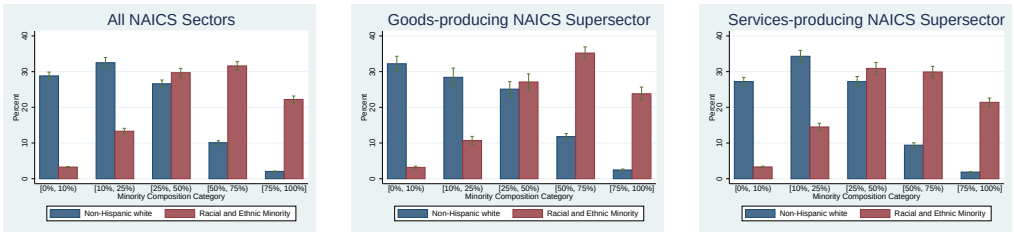
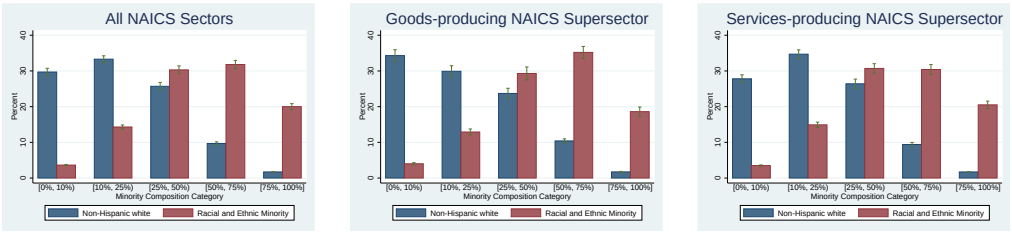


Figure 1.5. White and Minority Workers in Firms Grouped by Minority Composition Category

(a) 2000



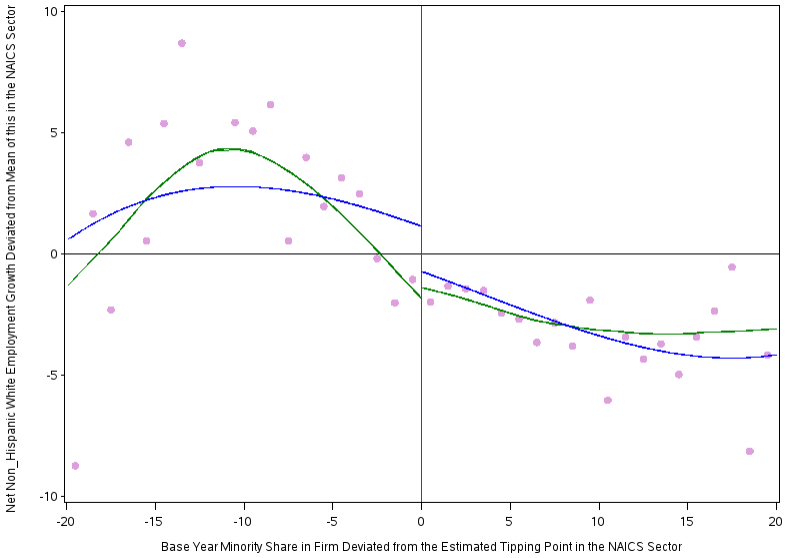
(b) 2005



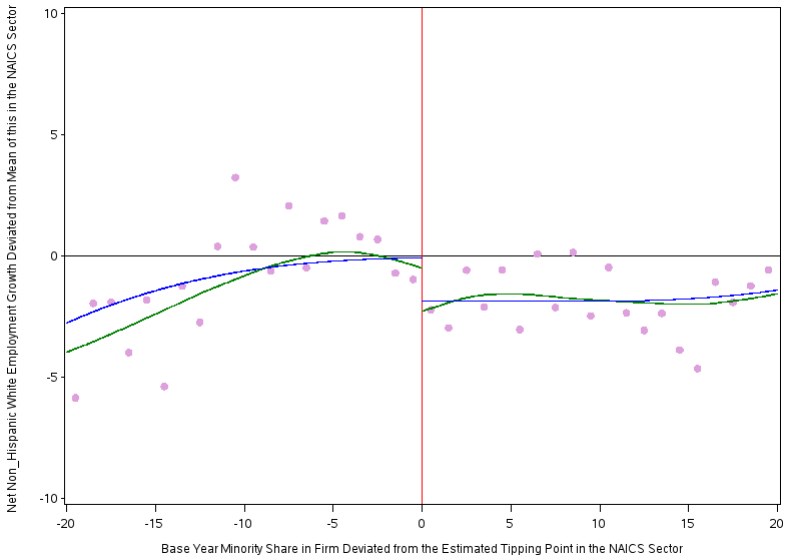
Notes: Blue bars represent non-Hispanic whites and red bars represents racial and ethnic minorities. Each of the statistics is computed and averaged across the results obtained using the 10 implicate files. The standard errors of the estimates are corrected taking into consideration of the variance contribution of multiple imputation. Standard error bars are included. The Rubin missingness ratios of these estimates can be found in Appendix Table 1.B.

Figure 1.6. Minority Composition Change in All NAICS Sectors Pooled Sample, by Relationship to Candidate Tipping Point

1995-2000



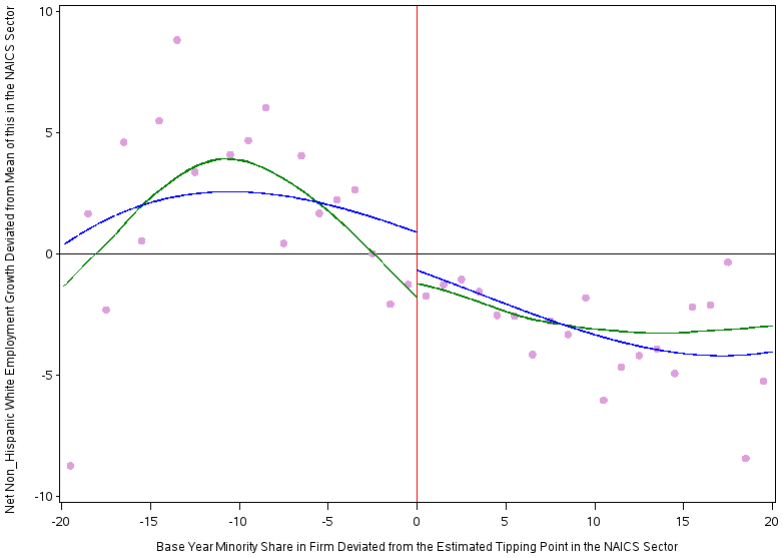
2000-2005



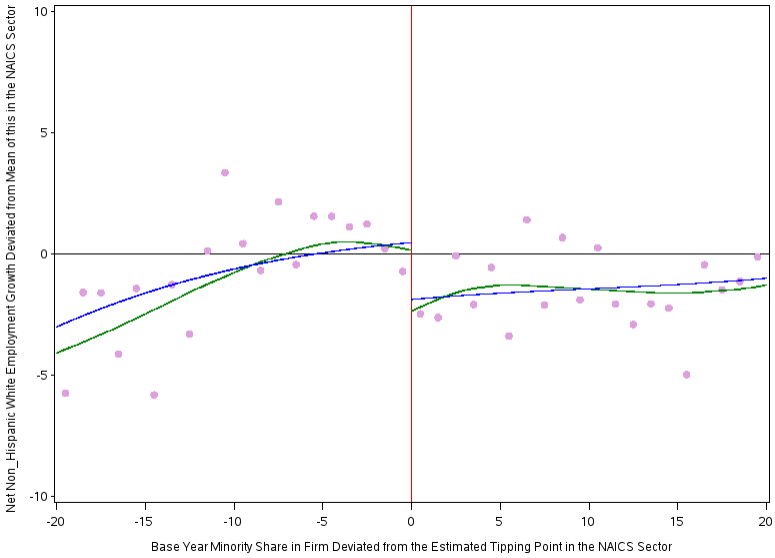
Notes: the X axis is minority share in establishment deviated from the estimated sector-specific tipping point. The Y axis is the change in net white employment in a five-year interval as a fraction of the total base year employment and deviated from the mean in the NAICS sector. Dots depict means in 1-percentage-point bins. The green line is a local linear regression fit separately on either side of zero using an Epanechnikov kernel and a bandwidth of 5. The blue line is a global third-order polynomial with an intercept shift at zero. All series use only the 50% of establishments not used to identify the tipping points.

Figure 1.7. Minority Composition Change in Service-producing NAICS Supersector Pooled Sample, by Relationship to Candidate Tipping Point

1995-2000



2000-2005



Notes: the X axis is minority share in establishment deviated from the estimated sector-specific tipping point. The Y axis is the change in net white employment in a five-year interval as a fraction of the total base year employment and deviated from the mean of this in the NAICS sector. Dots depict means in 1-percentage-point bins. The green line is a local linear regression fit separately on either side of zero using an Epanechnikov kernel and a bandwidth of 5. The blue line is a global 3^{rd} order polynomial with an intercept shift at zero. All series use only the 50% of establishments not used to identify the tipping points.

Table 1.1. Actual and Expected Duncan & Duncan Index by NAICS Sector

NAICS Sector	Duncan & Duncan (DD) Index		
	Actual	Expected	Difference
	Year 2000		
21 Mining, Quarrying, and Oil and Gas Extraction	36.8	13.4	23.3
22 Utilities	29.1	17.3	11.8
23 Construction	36.2	16.4	19.9
31-33 Manufacturing	41.1	11.1	30.0
42 Wholesale Trade	40.0	16.0	24.1
44-45 Retail Trade	40.2	17.0	23.2
48-49 Transportation and Warehousing	38.9	13.0	25.9
51 Information	31.8	10.7	21.1
52 Finance and Insurance	34.2	12.5	21.7
53 Real Estate and Rental and Leasing	37.9	17.4	20.5
54 Professional, Scientific, and Technical Services	31.7	15.1	16.6
55 Management of Companies and Enterprises	28.6	15.0	13.6
56 Administrative & Support and Waste Management & Remediation	40.5	11.2	29.3
61 Educational Services	35.3	11.3	24.0
62 Health Care and Social Assistance	42.7	12.1	30.6
71 Arts, Entertainment, and Recreation	35.2	12.4	22.8
72 Accommodation and Food Services	41.7	16.4	25.3
81 Other Services (except Public Administration)	42.7	18.4	24.3
	Year 2005		
21 Mining, Quarrying, and Oil and Gas Extraction	36.1	14.6	21.5
22 Utilities	31.6	17.6	13.9
23 Construction	35.9	17.6	18.3
31-33 Manufacturing	39.9	11.2	28.7
42 Wholesale Trade	38.3	16.1	22.3
44-45 Retail Trade	39.2	17.3	21.9
48-49 Transportation and Warehousing	38.4	13.1	25.3
51 Information	29.6	11.7	17.9
52 Finance and Insurance	32.9	13.1	19.8
53 Real Estate and Rental and Leasing	36.8	17.1	19.7
54 Professional, Scientific, and Technical Services	33.3	16.9	16.4
55 Management of Companies and Enterprises	28.8	12.3	16.5
56 Administrative & Support and Waste Management & Remediation	40.3	11.8	28.5
61 Educational Services	35.2	10.8	24.4
62 Health Care and Social Assistance	42.0	11.3	30.7
71 Arts, Entertainment, and Recreation	35.4	12.9	22.6
72 Accommodation and Food Services	40.5	16.0	24.5
81 Other Services (except Public Administration)	41.4	19.1	22.3

Notes: The sector-specific actual and expected DD indices are computed by averaging the state-specific actual and expected DD indices, weighted by the numbers of firms in the sector and state cell. Each statistic is computed and averaged across the results obtained using the 10 implicate files. The corrected standard errors and Rubin missingness ratios are presented in Appendix Tables 1.D and 1.E.

Table 1.2. Summary Statistics for Establishments

	1995			2000		
	All	**Goods-producing**	**Services-producing**	**All**	**Goods-producing**	**Services-producing**
Total # of Firms	200,000	48,500	151,000	341,000	78,300	263,000
Mean % Minority	33.60	34.40	33.40	33.20	32.90	33.30
Std. Dev.	(23.2)	(24.1)	(22.9)	(23.1)	(23.3)	(23.0)
Mean % Asians	5.88	4.83	6.21	5.95	4.51	6.38
Std. Dev.	(13.0)	(11.2)	(13.5)	(13.3)	(10.8)	(13.9)
Mean % Blacks	9.02	7.64	9.46	10.20	8.68	10.60
Std. Dev.	(14.9)	(12.7)	(15.5)	(15.7)	(13.5)	(16.3)
Mean % Hispanics	18.70	21.90	17.70	16.90	19.50	16.10
Std. Dev.	(21.4)	(23.1)	(20.7)	(20.5)	(21.9)	(20.0)
Growth in:						
White Employment	4.14	4.05	4.18	0.39	−3.77	2.17
Minority Employment	7.23	7.97	6.89	3.10	1.16	3.93
Asians	1.49	1.80	1.35	0.85	0.43	1.03
Blacks	1.73	1.00	2.05	0.46	−0.61	0.92
Hispanics	4.15	5.31	3.63	1.90	1.44	2.09
Total Employment	11.40	12.00	11.10	3.49	−2.60	6.10

Notes: Year at top of column is the base year. The numbers of firms do not sum up due to rounding for disclosure avoidance purposes. Each statistic is computed and averaged across the results obtained using the 10 implicate files. The corrected standard errors and Rubin missingness ratios are presented in Appendix Tables 1.F and 1.G.

Table 1.3. Summary Statistics for Establishments by Base-year Minority Shares

	1995			2000		
	All	**Goods-producing**	**Services-producing**	**All**	**Goods-producing**	**Services-producing**
Total # of Firms	200,000	48,500	151,000	341,000	78,300	263,000
0 to 5% Minority in BY:						
# of Firms	10,800	3,470	7,380	18,600	5,590	13,000
as % of Total # of Firms	5.40	7.15	4.89	5.45	7.14	4.94
Growth in:						
Total Employment	12.30	11.10	13.00	4.16	−3.75	9.07
White Employment	8.86	7.88	9.48	1.82	−5.29	6.23
5 to 20% Minority in BY:						
# of Firms	65,800	14,900	50,900	114,000	25,400	89,000
as % of Total # of Firms	32.90	30.72	33.71	33.43	32.44	33.84
Growth in:						
Total Employment	13.10	12.40	13.40	4.98	−1.88	7.76
White Employment	6.08	5.28	6.39	0.97	−4.70	3.26
20 to 50% Minority in BY:						
# of Firms	81,900	18,500	63,400	139,000	30,300	108,000
as % of Total # of Firms	40.95	38.14	41.99	40.76	38.70	41.06
Growth in:						
Total Employment	10.50	13.90	9.22	3.15	−1.80	5.02
White Employment	1.57	2.37	1.26	−0.92	−4.28	0.36
50 to 80% Minority in BY:						
# of Firms	33,200	9,520	23,700	55,600	14,200	41,400
as % of Total # of Firms	16.60	19.63	15.70	16.30	18.14	15.74
Growth in:						
Total Employment	9.65	11.70	8.48	1.01	−3.22	3.09
White Employment	1.85	1.91	1.81	0.24	−1.11	0.91
80 to 100% Minority in BY:						
# of Firms	8,170	2,160	6,010	13,500	2,850	10,700
as % of Total # of Firms	4.09	4.45	3.98	3.96	3.64	4.07
Growth in:						
Total Employment	7.90	3.49	10.20	1.68	−7.56	5.26
White Employment	3.36	2.69	3.70	2.64	1.63	3.03

Notes: "BY" stands for "Base Year." Year at the top of the column is the base year. The numbers of firms do not sum up due to rounding for disclosure avoidance purposes. Each statistic, except the number of firms as a percentage of the total number of firms, is computed and averaged across the results obtained using the 10 implicate files. The corrected standard errors and Rubin missingness ratios are presented in Appendix Table 1.H.

Table 1.4. NAICS Sector-Specific Candidate Tipping Points Using the Fixed-point Procedure

NAICS Sector	**Estimated Tipping Point**	
	$1995-2000$	$2000-2005$
21 Mining, Quarrying, and Oil and Gas Extraction	15.10	23.30
22 Utilities	10.90	18.60
23 Construction	14.20	9.74
31-33 Manufacturing	16.00	38.60
42 Wholesale Trade	13.20	7.18
44-45 Retail Trade	7.55	2.44
48-49 Transportation and Warehousing	19.20	9.90
51 Information	14.70	19.50
52 Finance and Insurance	12.50	13.00
53 Real Estate and Rental and Leasing	8.05	5.56
54 Professional, Scientific, and Technical Services	6.47	8.88
55 Management of Companies and Enterprises	15.80	16.80
56 Administrative & Support and Waste Management & Remediation	15.00	7.81
61 Educational Services	11.10	17.50
62 Health Care and Social Assistance	11.60	12.80
71 Arts, Entertainment, and Recreation	18.60	27.30
72 Accommodation and Food Services	39.70	26.20
81 Other Services (except Public Administration)	5.26	14.10
All NAICS Sector Average	14.16	15.51
Standard Deviation	7.49	9.08

Notes: Observations used to conduct the fixed-point procedure are the 50 percent simple random subsample of the establishments for each five-year interval. The tipping point is measured in base-year minority shares in each sector. Each estimate is computed and averaged across the results obtained using the 10 implicate files.

Table 1.5. Basic Regression Discontinuity Models for Changes in Employment Around the Tipping Point

	All NAICS						
	Net Change in white employment		Change in minority share	Change in white employment		Change in minority employment	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1995-2000							
Beyond candidate tipping point in 1995	−6.06	−5.83	2.32	−3.36	−3.15	2.70	2.68
	(1.14)	(1.16)	(0.33)	(1.02)	(1.04)	(0.54)	(0.52)
	[0.60]	[0.61]	[0.56]	[0.51]	[0.54]	[0.48]	[0.46]
Establishment controls	n	y	y	n	y	n	y
N	99,900	99,900	99,900	99,900	99,900	99,900	99,900
2000-2005							
Beyond candidate tipping point in 2000	−3.07	−3.25	0.96	−2.50	−2.70	0.57	0.56
	(1.06)	(1.06)	(0.45)	(1.01)	(1.03)	(0.59)	(0.59)
	[0.50]	[0.50]	[0.69]	[0.60]	[0.61]	[0.48]	[0.49]
Establishment controls	n	y	y	n	y	n	y
N	170,000	170,000	170,000	170,000	170,000	170,000	170,000

Notes: The unit of analysis is an establishment in the indicated five-year window. Dependent variables are the change in the relevant employment - net non-Hispanic white in columns (1) and (2), non-Hispanic white in columns (4) and (5), and minority in columns (6) and (7)) as a percentage (0-100) of the establishment's total base-year employment. Column (3) takes as the dependent variable the minority share in an end year minus the minority share in a base year but is otherwise identical. All specifications are estimated using only the 50% of single-establishment firms not used to identify the candidate tipping points. The specifications include fixed state effects, fixed NAICS sector effects, a cubic polynomial in the deviation in the establishment's minority share from the candidate tipping point, the share of workers approaching retirement, the share of workers who are 24 years old or younger, and log average earnings. All are measured in the base year at the establishment-level. Standard errors are clustered on the state-sector level. All estimates are computed and averaged across the results obtained using the 10 implicate files. The variance-covariance matrices of the estimates are corrected, taking into consideration the variance contribution of multiple imputation. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.6. Basic Regression Discontinuity Models for Changes in Employment Around the Tipping Point Using the 19 States Contained in 1995-2000 Sample, 2000-2005

	All NAICS			
	Net Change in white employment	Change in minority share	Change in white employment	Change in minority employment
	(1)	(2)	(3)	(4)
2000-2005				
Beyond candidate TP in 2000	−2.75	0.86	−2.29	0.46
	(1.20)	(0.54)	(1.08)	(0.73)
	[0.34]	[0.59]	[0.46]	[0.42]
Establishment controls	y	y	y	y
N	119,000	119,000	119,000	119,000

Notes: The unit of analysis is an establishment in 2000-2005 from the following 19 states: CA, CO, FL, ID, IL, KS, LA, MD, MN, MO, MT, NC, NY, OR, PA, RI, TX, WA, and WI. See Table 1.5 footnote (columns(2), (3), (5), and (7)) for details on specifications. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.7. Basic Regression Discontinuity Models for Net Non-Hispanic White Employment Changes Around the Tipping Point: Goods-producing vs. Services-producing NAICS Supersectors

	Services-producing			**Goods-producing**		
	Net Change in white employment	Change in white employment	Change in minority employment	Net Change in white employment	Change in white employment	Change in minority employment
	(1)	(2)	(3)	(4)	(5)	(6)
1995-2000						
Beyond candidate tipping	−5.78	−3.32	2.46	−0.14	−0.067	0.068
point in 1995	(1.18)	(1.05)	(0.51)	(3.34)	(3.09)	(1.99)
	[0.59]	[0.50]	[0.42]	[0.49]	[0.28]	[0.55]
Establishment controls	y	y	y	y	y	y
N	90,100	90,100	90,100	9,800	9,800	9,800
2000-2005						
Beyond candidate tipping	−3.34	−2.45	0.89	−0.16	−0.23	−0.067
point in 2000	(1.11)	(1.07)	(0.59)	(4.52)	(4.32)	(2.32)
	[0.50]	[0.63]	[0.53]	[0.51]	[0.45]	[0.66]
Establishment controls	y	y	y	y	y	y
N	152,000	152,000	152,000	18,000	18,000	18,000

Notes: See Table 1.5 footnote for a description on specifications. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.8. Tipping in Firms Which Have Undergone Small Changes in Employment

	All NAICS				
	Net Change in white employment	Change in minority share	Change in white employment	Change in minority employment	Change in total employment
	(1)	(2)	(3)	(4)	(5)
1995-2000					
Beyond candidate tipping point in 1995	−4.24	2.05	−2.18	2.06	−0.11
	(0.82)	(0.41)	(0.43)	(0.41)	(0.19)
	[0.44]	[0.45]	[0.42]	[0.46]	[0.50]
Establishment controls	y	y	y	y	y
N	21,500	21,500	21,500	21,500	21,500
2000-2005					
Beyond candidate tipping point in 2000	−1.87	0.85	−1.01	0.86	−0.16
	(0.92)	(0.45)	(0.47)	(0.45)	(0.14)
	[0.54]	[0.56]	[0.52]	[0.55]	[0.35]
Establishment controls	y	y	y	y	y
N	39,100	39,100	39,100	39,100	39,100

Notes: The unit of analysis is an establishment in a five-year interval that has undergone less than 10 percentage point change in establishment-level employment. See Table 1.5 footnote (columns(2), (3), (5), and (7)) for details on specifications. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.9. Sensitivity to Flexible Controls For Establishment Covariates

	(1)	(2)	(3)	(4)	(5)
1995-2000	−5.83	−5.85	−6.02	−5.99	−6.11
	(1.16)	(1.17)	(1.18)	(1.18)	(1.20)
	[0.61]	[0.62]	[0.63]	[0.63]	[0.64]
3^{rd}-order polynomial in:					
log average earnings		y			y
share of retiring workers			y		y
share of young workers				y	y
2000-2005	−3.25	−3.26	−3.22	−3.17	−3.17
	(1.06)	(1.07)	(1.08)	(1.08)	(1.10)
	[0.50]	[0.50]	[0.49]	[0.50]	[0.50]

Notes: The specification in column (1) is that from column (2) of Table 1.5. The dependent variable is the change in net non-Hispanic white employment as a percentage of base-year total establishment employment. The remaining specifications add third-order polynomials in the listed control variables. All specifications are estimated using only the 50% of establishments not used to identify the tipping points. Standard errors are clustered on the state-sector level. All estimates are computed and averaged across the results obtained using the 10 implicate files. The variance-covariance matrices of the estimates are corrected, taking into consideration the variance contribution of multiple imputation. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.10. Changes in Covariates Around the Candidate Tipping Point

Dependent Variable: Change in	Log Average Earnings (1)	% of Retiring Workers (2)	% of Young Workers (3)
1995-2000			
Beyond candidate tipping point in 1995	0.0019	0.049	−0.032
	(0.0057)	(0.21)	(0.26)
	[0.38]	[0.57]	[0.53]
2000-2005			
Beyond candidate tipping point in 2000	−0.00052	−0.052	−0.19
	(0.0048)	(0.15)	(0.21)
	[0.27]	[0.30]	[0.49]

Notes: All specifications are estimated using only the 50% of single-establishment firms not used to identify the candidate tipping points. All specifications include fixed state effects, fixed sector effects, and a cubic polynomial in the deviation in the establishment's minority share from the candidate tipping point. Standard errors are clustered on the state-NAICS sector. All estimates are computed and averaged across the results obtained using the 10 implicate files. The variance-covariance matrices of the estimates are corrected, taking into consideration the variance contribution of multiple imputation. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.11. Tipping in Minority Share, Black Share, and Hispanic Share, 1995-2000

	Percentage change in net white employment				Percentage change in white employment				Percentage change in minority employment			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Beyond minority share fixed point	−5.83			−4.95	−3.15			−2.03	2.68			2.91
	(1.16)			(1.14)	(1.04)			(1.10)	(0.52)			(0.55)
	[0.61]			[0.60]	[0.54]			[0.56]	[0.46]			[0.42]
Beyond black share fixed point		−4.68		−4.20		−7.71		−7.18		−3.03		−2.98
		(1.10)		(1.16)		(1.71)		(1.79)		(1.46)		(1.43)
		[0.55]		[0.60]		[0.60]		[0.62]		[0.63]		[0.66]
Beyond Hispanic share fixed point			−4.71	−3.15			−6.56	−5.54			−1.84	−2.39
			(0.79)	(0.73)			(1.01)	(0.95)			(0.78)	(0.75)
			[0.27]	[0.22]			[0.16]	[0.24]			[0.23]	[0.24]
Cubic in min. share minus TP	y			y	y			y	y			y
Cubic in bl. share minus TP		y		y		y		y		y		y
Cubic in hi. share minus TP			y	y			y	y			y	y
Establishment controls	y	y	y	y	y	y	y	y	y	y	y	y

Notes: See Table 1.5 footnote for a description of establishment-level controls. Specification in columns (1), (5), and (9) are identical to those in Table 1.5. Other columns explore candidate tipping points in the establishment black share or Hispanic share. All specifications include fixed state effects and fixed sector effects. All specifications are estimated using only the 50% of establishments not used to identify the tipping points. Standard errors are clustered on the state-sector level. All estimates are computed and averaged across the results obtained using the 10 implicate files. The variance-covariance matrices of the estimates are corrected, taking into consideration the variance contribution of multiple imputation. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Table 1.12. Tipping in Minority Share, Black Share, and Hispanic Share, 2000-2005

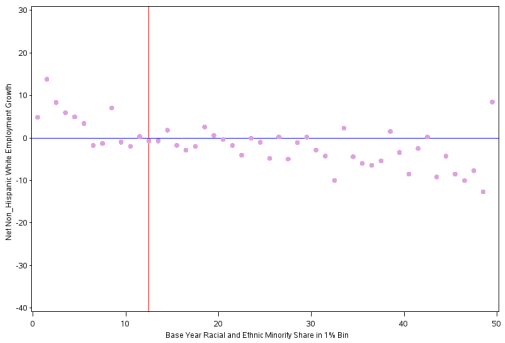
	net white employment				**Percentage change in white employment**				**minority employment**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Beyond minority share fixed point	−3.25			−2.39	−2.70			−1.66	0.56			0.72
	(1.06)			(1.00)	(1.03)			(1.02)	(0.59)			(0.66)
	[0.50]			[0.46]	[0.61]			[0.61]	[0.49]			[0.57]
Beyond black share fixed point		0.29		0.039		0.84		0.52		0.55		0.48
		(1.18)		(0.95)		(2.00)		(1.76)		(1.12)		(1.21)
		[0.69]		[0.53]		[0.85]		[0.82]		[0.75]		[0.79]
Beyond Hispanic share fixed point			−4.54	−2.89			−7.01	−5.84			−2.47	−2.94
			(0.74)	(0.71)			(1.24)	(1.18)			(0.98)	(0.99)
			[0.39]	[0.42]			[0.69]	[0.72]			[0.75]	[0.75]
Cubic in min. share minus TP	y			y	y			y	y			y
Cubic in bl. share minus TP		y		y		y		y		y		y
Cubic in hi. share minus TP			y	y			y	y			y	y
Establishment controls	y	y	y	y	y	y	y	y	y	y	y	y

Notes: See Table 1.5 footnote for a description of establishment-level controls. Specification in columns (1), (5) and (9) are identical to those in Table 1.5. Other columns explore candidate tipping points in the establishment black share or Hispanic share. All specifications include fixed state effects and fixed sector effects. All specifications are estimated using only the 50% of establishments not used to identify the tipping points. Standard errors are clustered on the state-sector level. All estimates are computed and averaged across the results obtained using the 10 implicate files. The variance-covariance matrices of the estimates are corrected, taking into consideration the variance contribution of multiple imputation. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

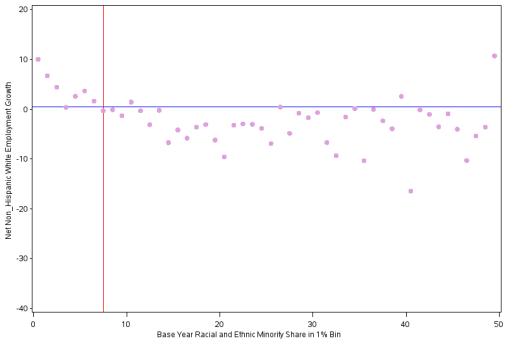
Appendix Figure 1.A. Firm-level Minority Composition Change in Selected NAICS Sectors

(a) 1995-2000

NAICS 42: Wholesale Trade

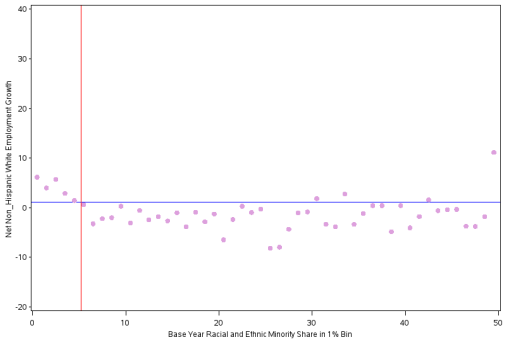


NAICS 44-45: Retail Trade

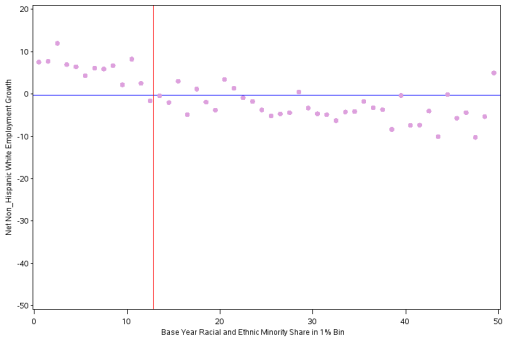


(b) 2000-2005

NAICS 42: Wholesale Trade



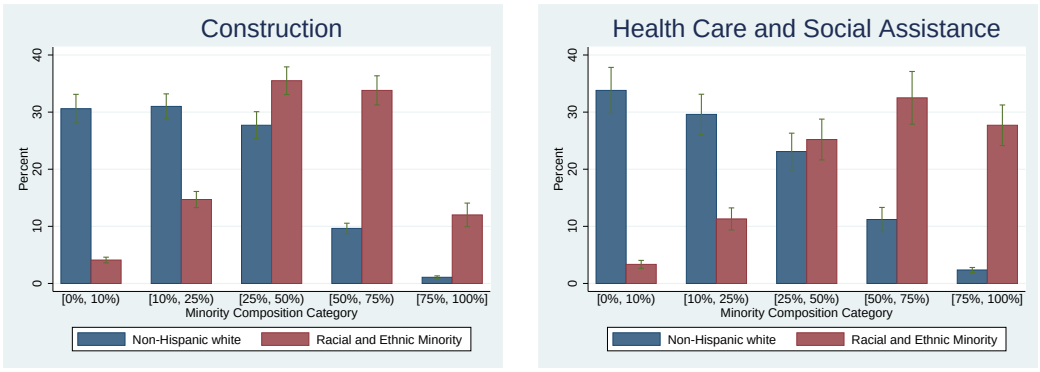
NAICS 62: Health Care and Social Assistance



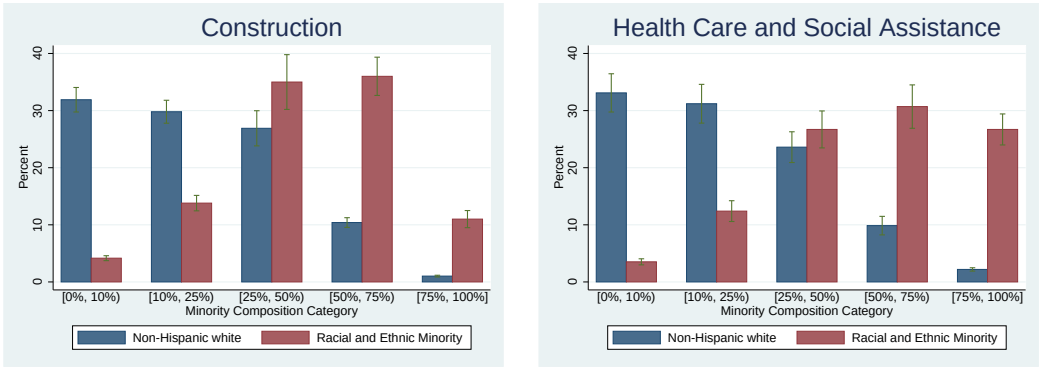
Notes: Dots show mean of the change in the net establishment-level white employment in a five-year window as a percentage of the total employment in the base year, grouping establishments into cells of width 1 percentage point by the base-year minority share. The horizontal line depicts the unconditional mean. The vertical line depicts the estimated tipping point using the fixed-point procedure described in subsection 1.2.3 and a 50 percent sample of single-establishment firms in a NAICS sector.

Appendix Figure 1.B. White and Minority Workers in Firms Grouped by Minority Composition Category in Selected NAICS Sectors

(a) 2000



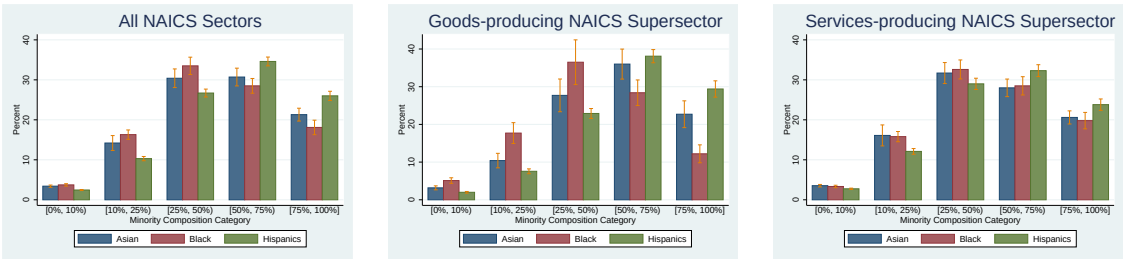
(b) 2005



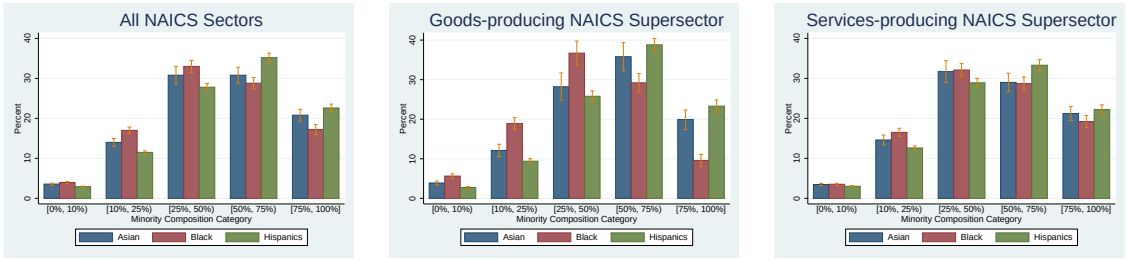
Notes: Blue bars represent non-Hispanic whites and red bars represents racial and ethnic minorities. Each of the statistics is computed and averaged across the results obtained using the 10 implicate files. The standard errors of the estimates are corrected taking into consideration of the variance contribution of multiple imputation. Standard error bars are included. The Rubin missingness ratios of these estimates can be found in Appendix Table 1.C.

Appendix Figure 1.C. Various Groups of Minority Workers in Firms Grouped by Minority Composition Category

(a) 2000



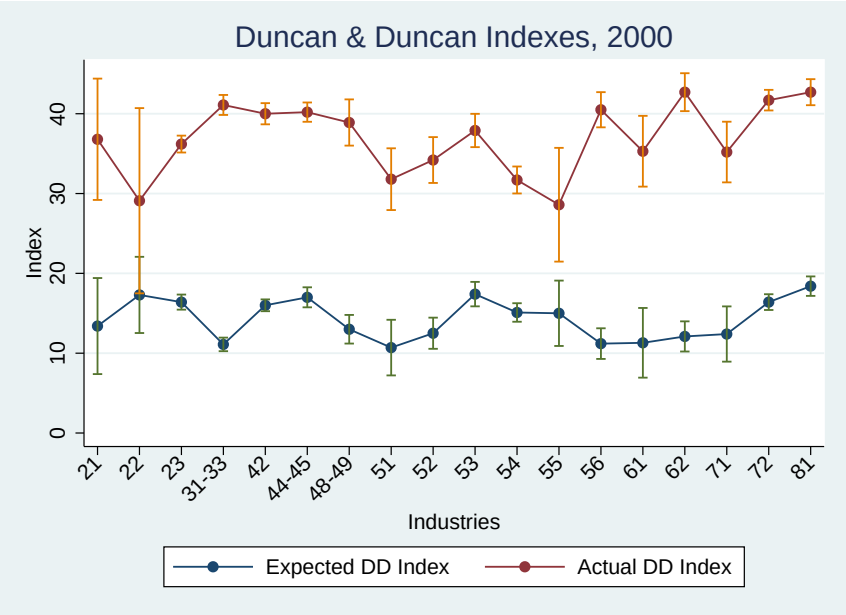
(b) 2005



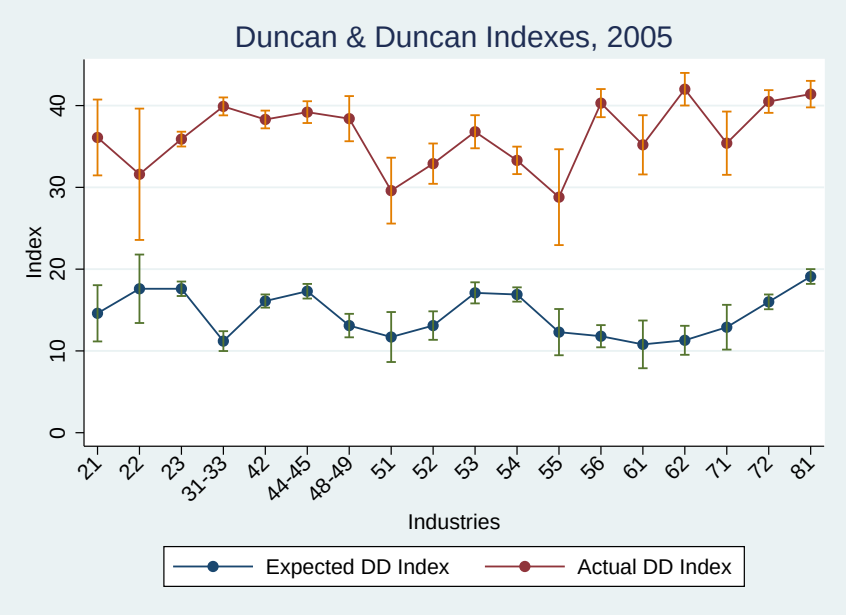
Notes: Blue bars represent Asians, red bars represent blacks, and green bars represent Hispanics. Each of the statistics is computed and averaged across the results obtained using the 10 implicate files. The standard errors of the estimates are corrected taking into consideration of the variance contribution of multiple imputation. Standard error bars are included. The Rubin missingness ratios of these estimates can be found in Appendix Table 1.B.

Appendix Figure 1.D. Expected and Actual Duncan & Duncan Index By Industry

2000

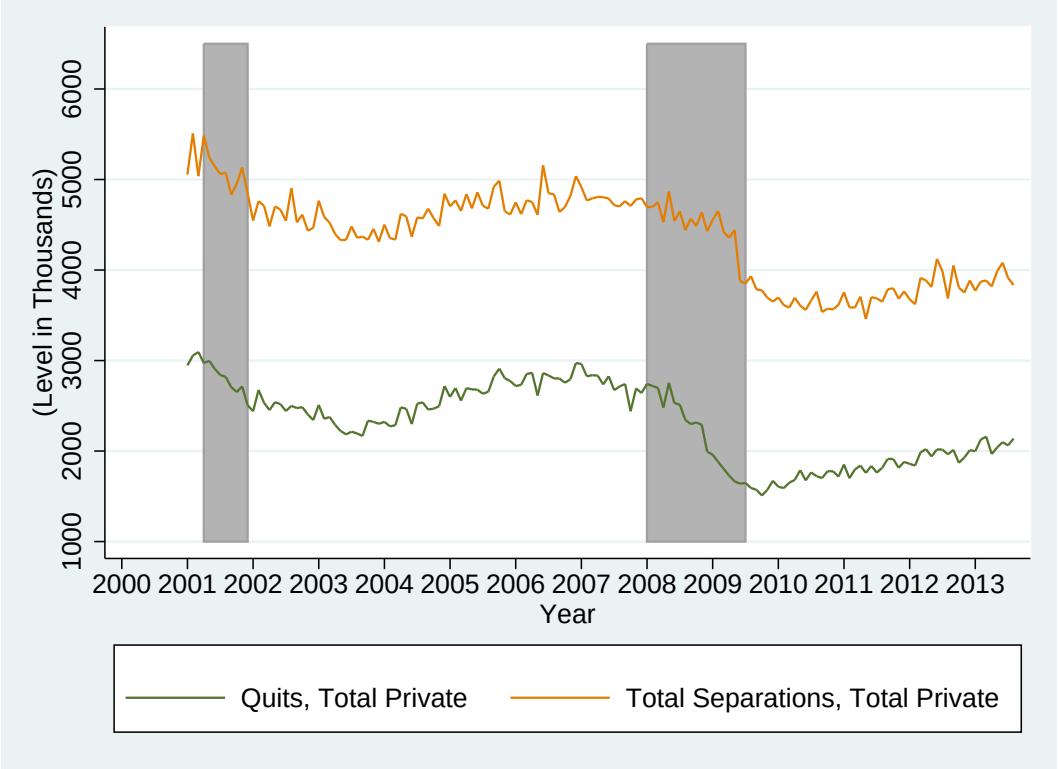


2005



Notes: Blue dots represent the expected Duncan & Duncan index, red dots represent the actual Duncan & Duncan index. Each of the statistics is computed and averaged across the results obtained using the 10 implicate files. The standard errors of the estimates are corrected taking into consideration of the variance contribution of multiple imputation. Standard error bars are included, which are taken from Appendix Table 1.D and Appendix Table 1.E. The Rubin missingness ratios of these estimates can be found in Appendix Table 1.D and Appendix Table 1.E.

Appendix Figure 1.E. Quits and Total Separations: Total Private, Monthly, Seasonally Adjusted, 2000-12-01 to 2013-07-01



Data Source: Bureau of Labor Statistics, Job Openings and Labor Turnover Survey, JTS1000QUL and JTS1000TSL. Shaded areas indicate U.S. recessions. The start and end dates of the recessions are obtained from the National Bureau of Economic Research, (`http://www.nber.org/cycles.html`)

Appendix Table 1.A. Summary Statistics for Establishments by Base-year Minority Shares of Selected NAICS Sectors

	1995		2000	
	23 Construction	**62 Health Care & Social Asst**	**23 Construction**	**62 Health Care & Social Asst**
Total # of Firms	19,000	22,500	35,400	42,000
Mean % Minority	31.30	34.90	31.20	34.30
Std. Dev.	(21.9)	(23.3)	(21.6)	(23.7)
Growth in:				
White Employment	14.60	2.53	2.16	5.93
Minority Employment	11.80	5.42	5.52	6.93
Total Employment	26.40	7.95	7.68	12.90
0 to 5% Minority in BY:				
# of Firms	1,140	1,110	2,070	2,170
as % of Total # of Firms	6.00	4.93	5.85	5.17
Growth in:				
Total Employment	27.10	10.30	3.81	14.20
White Employment	22.10	8.18	1.17	12.00
5 to 20% Minority in BY:				
# of Firms	6,750	6,970	12,400	13,700
as % of Total # of Firms	35.53	30.98	35.03	32.62
Growth in:				
Total Employment	27.90	9.57	5.81	14.30
White Employment	17.10	3.96	0.63	9.11
20 to 50% Minority in BY:				
# of Firms	7,880	9,710	14,800	17,200
as % of Total # of Firms	41.47	43.16	41.81	40.95
Growth in:				
Total Employment	25.40	7.64	9.06	10.90
White Employment	10.40	−0.20	2.15	2.99
50 to 80% Minority in BY:				
# of Firms	2,750	3,720	5,320	6,860
as % of Total # of Firms	14.47	16.53	15.03	16.33
Growth in:				
Total Employment	24.90	4.49	11.50	11.30
White Employment	10.90	0.33	6.03	1.40
80 to 100% Minority in BY:				
# of Firms	500	1,030	810	2,020
as % of Total # of Firms	2.63	4.58	2.29	4.81
Growth in:				
Total Employment	21.80	6.03	14.50	15.90
White Employment	10.50	1.33	9.14	2.27

Notes: See Table 1.3 footnote for a description of the table structure. The corrected standard errors and Rubin missingness ratios are presented in Appendix Tables 1.I and 1.J.

Appendix Table 1.B. Rubin Missingness Ratios Computed for Estimates Used in Figure 1.5 and Appendix Figures 1.C

NAICS Sector	Minority Composition	Percent of				
		Whites	**Minorities**	**Asians**	**Blacks**	**Hispanics**
		Year 2000				
ALL	[0, 10%)	0.25	0.25	0.29	0.34	0.26
	[10%, 25%)	0.37	0.38	0.26	0.32	0.47
	[25%, 50%)	0.18	0.26	0.36	0.34	0.28
	[50%, 75%)	0.35	0.34	0.48	0.26	0.24
	[75%, 100%]	0.28	0.29	0.22	0.56	0.31
Goods-producing Super-sector	[0, 10%)	0.27	0.49	0.46	0.39	0.50
	[10%, 25%)	0.27	0.39	0.49	0.31	0.42
	[25%, 50%)	0.24	0.28	0.23	0.30	0.31
	[50%, 75%)	0.16	0.073	0.36	0.23	0.20
	[75%, 100%]	0.21	0.29	0.23	0.31	0.37
Services-producing Super-sector	[0, 10%)	0.19	0.20	0.12	0.29	0.33
	[10%, 25%)	0.33	0.33	0.26	0.32	0.40
	[25%, 50%)	0.40	0.44	0.46	0.52	0.31
	[50%, 75%)	0.30	0.32	0.32	0.31	0.30
	[75%, 100%]	0.26	0.31	0.24	0.50	0.40
		Year 2005				
ALL	[0, 10%)	0.61	0.63	0.40	0.60	0.62
	[10%, 25%)	0.35	0.35	0.45	0.28	0.30
	[25%, 50%)	0.51	0.41	0.25	0.43	0.38
	[50%, 75%)	0.32	0.37	0.34	0.26	0.37
	[75%, 100%]	0.29	0.35	0.40	0.35	0.29
Goods-producing Super-sector	[0, 10%)	0.47	0.52	0.40	0.52	0.51
	[10%, 25%)	0.30	0.24	0.29	0.20	0.34
	[25%, 50%)	0.33	0.20	0.26	0.24	0.22
	[50%, 75%)	0.25	0.30	0.44	0.25	0.28
	[75%, 100%]	0.33	0.25	0.17	0.29	0.23
Services-producing Super-sector	[0, 10%)	0.52	0.51	0.32	0.46	0.54
	[10%, 25%)	0.41	0.44	0.51	0.36	0.32
	[25%, 50%)	0.44	0.41	0.24	0.45	0.33
	[50%, 75%)	0.30	0.35	0.26	0.25	0.37
	[75%, 100%]	0.32	0.35	0.43	0.34	0.30

Notes: The variable "percent of whites" is used to create the blue bars in Figure 1.5; the variable "percent of minorities" is used to create the red bars in Figure 1.5; the variable "percent of Asians" is used to create the blue bars in Appendix Figure 1.C; the variable "percent of blacks" is used to create the red bars in Appendix Figure 1.C; the variable "percent of Hispanics" is used to create the green bars in Appendix Table 1.C. The Rubin missingness ratios are computed following the computational formulas presented in Appendix 1.7.D.

Appendix Table 1.C. Rubin Missingness Ratios Computed for Estimates Used in Appendix Figures 1.B

NAICS Sector	Minority Composition	Percent of Whites	Percent of Minorities
		Year 2000	
23 Construction	[0, 10%)	0.47	0.46
	[10%, 25%)	0.30	0.38
	[25%, 50%)	0.33	0.19
	[50%, 75%)	0.17	0.25
	[75%, 100%]	0.25	0.28
62 Health Care & Social Assistance	[0, 10%)	0.28	0.35
	[10%, 25%)	0.12	0.20
	[25%, 50%)	0.46	0.53
	[50%, 75%)	0.17	0.30
	[75%, 100%]	0.24	0.29
		Year 2005	
23 Construction	[0, 10%)	0.44	0.39
	[10%, 25%)	0.37	0.25
	[25%, 50%)	0.23	0.21
	[50%, 75%)	0.28	0.33
	[75%, 100%]	0.15	0.19
62 Health Care & Social Assistance	[0, 10%)	0.50	0.48
	[10%, 25%)	0.43	0.35
	[25%, 50%)	0.48	0.53
	[50%, 75%)	0.31	0.39
	[75%, 100%]	0.14	0.24

Notes: The variable "percent of whites" is used to create the blue bars in Appendix Figure 1.B; the variable "percent of minorities" is used to create the red bars in Figure 1.B. The Rubin missingness ratios are computed following the computational formulas presented in Appendix 1.7.D.

Appendix Table 1.D. The Corrected Standard Errors and Rubin Missingness Ratios Computed for the Actual and Expected Duncan & Duncan Index in Table 1.1, Year 2000

NAICS Sector	Duncan & Duncan Index		
	Actual	**Expected**	**Difference**
21 Mining, Quarrying, and Oil and Gas Extraction	(3.36) [0.23]	(2.66) [0.21]	(3.39) [0.33]
22 Utilities	(5.13) [0.23]	(2.11) [0.19]	(4.45) [0.15]
23 Construction	(0.47) [0.19]	(0.42) [0.15]	(0.60) [0.21]
31-33 Manufacturing	(0.55) [0.21]	(0.38) [0.22]	(0.64) [0.23]
42 Wholesale Trade	(0.59) [0.21]	(0.33) [0.51]	(0.67) [0.38]
44-45 Retail Trade	(0.53) [0.18]	(0.55) [0.36]	(0.80) [0.16]
48-49 Transportation and Warehousing	(1.28) [0.44]	(0.79) [0.39]	(1.08) [0.35]
51 Information	(1.71) [0.16]	(1.54) [0.13]	(2.37) [0.15]
52 Finance and Insurance	(1.27) [0.26]	(0.86) [0.19]	(1.44) [0.29]
53 Real Estate and Rental and Leasing	(0.92) [0.17]	(0.68) [0.32]	(0.97) [0.25]
54 Professional, Scientific, and Technical Services	(0.75) [0.17]	(0.51) [0.36]	(0.73) [0.22]
55 Management of Companies and Enterprises	(3.15) [0.20]	(1.81) [0.35]	(3.47) [0.25]
56 Administrative & Support and Waste Management & Remediation	(0.97) [0.29]	(0.85) [0.36]	(1.09) [0.27]
61 Educational Services	(1.96) [0.20]	(1.93) [0.21]	(2.32) [0.20]
62 Health Care and Social Assistance	(1.05) [0.19]	(0.84) [0.28]	(1.09) [0.18]
71 Arts, Entertainment, and Recreation	(1.68) [0.26]	(1.53) [0.37]	(2.46) [0.38]
72 Accommodation and Food Services	(0.57) [0.11]	(0.44) [0.55]	(0.68) [0.22]
81 Other Services (except Public Administration)	(0.72) [0.18]	(0.54) [0.37]	(0.86) [0.30]

Notes: The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Appendix Table 1.E. The Corrected Standard Errors and Rubin Missingness Ratios Computed for the Actual and Expected Duncan & Duncan Index in Table 1.1, Year 2005

NAICS Sector	Duncan & Duncan Index		
	Actual	**Expected**	**Difference**
21 Mining, Quarrying, and Oil and Gas Extraction	(2.05) [0.09]	(1.52) [0.15]	(2.37) [0.14]
22 Utilities	(3.55) [0.12]	(1.85) [0.53]	(4.03) [0.21]
23 Construction	(0.40) [0.17]	(0.39) [0.40]	(0.53) [0.46]
31-33 Manufacturing	(0.48) [0.19]	(0.54) [0.34]	(0.67) [0.38]
42 Wholesale Trade	(0.48) [0.29]	(0.36) [0.18]	(0.52) [0.11]
44-45 Retail Trade	(0.59) [0.22]	(0.40) [0.53]	(0.50) [0.23]
48-49 Transportation and Warehousing	(1.22) [0.44]	(0.63) [0.19]	(1.26) [0.30]
51 Information	(1.78) [0.14]	(1.35) [0.21]	(2.25) [0.23]
52 Finance and Insurance	(1.09) [0.36]	(0.77) [0.27]	(1.23) [0.23]
53 Real Estate and Rental and Leasing	(0.89) [0.20]	(0.57) [0.10]	(0.80) [0.21]
54 Professional, Scientific, and Technical Services	(0.74) [0.31]	(0.39) [0.33]	(0.75) [0.22]
55 Management of Companies and Enterprises	(2.59) [0.40]	(1.25) [0.33]	(2.62) [0.33]
56 Administrative & Support and Waste Management & Remediation	(0.76) [0.23]	(0.60) [0.29]	(0.80) [0.28]
61 Educational Services	(1.60) [0.31]	(1.29) [0.12]	(2.03) [0.23]
62 Health Care and Social Assistance	(0.88) [0.14]	(0.78) [0.11]	(0.93) [0.23]
71 Arts, Entertainment, and Recreation	(1.71) [0.078]	(1.21) [0.26]	(2.17) [0.17]
72 Accommodation and Food Services	(0.61) [0.43]	(0.40) [0.16]	(0.70) [0.38]
81 Other Services (except Public Administration)	(0.72) [0.23]	(0.40) [0.18]	(0.79) [0.27]

Notes: The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Appendix Table 1.F. The Corrected Standard Errors and Rubin Missingness Ratios for Summary Statistics in Table 1.2, Part 1

	1995			2000		
	All	**Goods-producing**	**Services-producing**	**All**	**Goods-producing**	**Services-producing**
Mean % Minority	(0.065)	(0.12)	(0.081)	(0.050)	(0.096)	(0.060)
	[0.37]	[0.22]	[0.48]	[0.36]	[0.22]	[0.44]
Std. Dev.	(0.040)	(0.074)	(0.053)	(0.031)	(0.060)	(0.039)
	[0.39]	[0.29]	[0.52]	[0.33]	[0.31]	[0.45]
Mean % Asians	(0.035)	(0.060)	(0.043)	(0.029)	(0.046)	(0.035)
	[0.32]	[0.27]	[0.35]	[0.40]	[0.28]	[0.40]
Std. Dev.	(0.063)	(0.13)	(0.07)	(0.060)	(0.11)	(0.069)
	[0.23]	[0.28]	[0.22]	[0.50]	[0.29]	[0.50]
Mean % Blacks	(0.036)	(0.064)	(0.045)	(0.032)	(0.052)	0.038
	[0.17]	[0.18]	[0.25]	[0.29]	[0.14]	[0.29]
Std. Dev.	(0.055)	(0.10)	(0.066)	(0.050)	(0.083)	(0.058)
	[0.25]	[0.23]	[0.28]	[0.48]	[0.30]	[0.48]
Mean % Hispanics	(0.059)	(0.12)	(0.069)	(0.039)	(0.090)	(0.048)
	[0.35]	[0.17]	[0.41]	[0.16]	[0.23]	[0.33]
Std. Dev.	(0.052)	(0.091)	(0.066)	(0.037)	(0.075)	(0.045)
	[0.42]	[0.30]	[0.49]	[0.23]	[0.30]	[0.31]

Notes: The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Appendix Table 1.G. The Corrected Standard Errors and Rubin Missingness Ratios for Summary Statistics in Table 1.2, Part 2

	1995			2000		
	All	**Goods-producing**	**Services-producing**	**All**	**Goods-producing**	**Services-producing**
Growth in:						
White Employment	(0.31)	(0.46)	(0.41)	(0.20)	(0.31)	(0.23)
	[0.28]	[0.25]	[0.31]	[0.27]	[0.24]	[0.22]
Minority Employment	(0.19)	(0.33)	(0.23)	(0.12)	(0.23)	(0.15)
	[0.28]	[0.26]	[0.22]	[0.23]	[0.41]	[0.27]
Asians	(0.069)	(0.14)	(0.074)	(0.043)	(0.089)	(0.052)
	[0.34]	[0.26]	[0.33]	[0.24]	[0.47]	[0.28]
Blacks	(0.096)	(0.13)	(0.12)	(0.059)	(0.098)	(0.078)
	[0.23]	[0.38]	[0.13]	[0.15]	[0.41]	[0.18]
Hispanics	(0.098)	(0.21)	(0.12)	(0.069)	(0.14)	(0.080)
	[0.28]	[0.24]	[0.42]	[0.42]	[0.42]	[0.41]
Total Employment	(0.46)	(0.67)	(0.58)	(0.28)	(0.46)	(0.33)
	[0.28]	[0.20]	[0.28]	[0.21]	[0.29]	[0.17]

Notes: The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Appendix Table 1.H. The Corrected Standard Errors and Rubin Missingness Ratios for Summary Statistics in Table 1.3

	1995			2000		
	All	**Goods-producing**	**Services-producing**	**All**	**Goods-producing**	**Services-producing**
0 to 5% Minority in BY:						
Growth in:						
Total Employment	(0.83)	(1.34)	(1.15)	(0.57)	(0.99)	(0.74)
	[0.35]	[0.35]	[0.46]	[0.28]	[0.32]	[0.42]
White Employment	(0.78)	(1.23)	(1.09)	(0.53)	(0.94)	(0.69)
	[0.37]	[0.33]	[0.47]	[0.26]	[0.33]	[0.41]
5 to 20% Minority in BY:						
Growth in:						
Total Employment	(0.82)	(1.39)	(0.99)	(0.51)	(0.90)	(0.59)
	[0.29]	[0.26]	[0.28]	[0.43]	[0.35]	[0.46]
White Employment	(0.66)	(1.11)	(0.81)	(0.42)	(0.73)	(0.49)
	[0.30]	[0.27]	[0.32]	[0.42]	[0.34]	[0.47]
20 to 50% Minority in BY:						
Growth in:						
Total Employment	(0.77)	(1.38)	(0.97)	(0.53)	(0.81)	(0.66)
	[0.10]	[0.25]	[0.15]	[0.19]	[0.21]	[0.18]
White Employment	(0.49)	(0.83)	(0.62)	(0.33)	(0.46)	(0.42)
	[0.11]	[0.27]	[0.15]	[0.13]	[0.12]	[0.13]
50 to 80% Minority in BY:						
Growth in:						
Total Employment	(1.26)	(1.67)	(1.72)	(0.73)	(1.13)	(1.01)
	[0.36]	[0.40]	[0.33]	[0.23]	[0.35]	[0.31]
White Employment	(0.52)	(0.59)	(0.72)	(0.29)	(0.41)	(0.40)
	[0.39]	[0.46]	[0.34]	[0.22]	[0.30]	[0.24]
80 to 100% Minority in BY:						
Growth in:						
Total Employment	(1.44)	(2.55)	(1.72)	(1.13)	(2.14)	(1.37)
	[0.27]	[0.25]	[0.28]	[0.27]	[0.35]	[0.28]
White Employment	(0.31)	(0.48)	(0.39)	(0.24)	(0.36)	(0.30)
	[0.35]	[0.32]	[0.32]	[0.36]	[0.16]	[0.39]

Notes: "BY" stands for "Base Year." Year at the top of the column is the base year. The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Appendix Table 1.I. The Corrected Standard Errors and Rubin Missingness Ratios for Summary Statistics in Appendix Table 1.A, Part 1

	1995		2000	
	23 Construction	**62 Health Care & Social Asst**	**23 Construction**	**62 Health Care & Social Asst**
Mean % Minority	(0.17)	(0.22)	(0.14)	(0.14)
	[0.16]	[0.47]	[0.28]	[0.28]
Std. Dev.	(0.12)	(0.12)	(0.097)	(0.077)
	[0.30]	[0.33]	[0.40]	[0.17]
Growth in:				
White Employment	(0.84)	(0.76)	(0.48)	(0.46)
	[0.11]	[0.27]	[0.35]	[0.26]
Minority Employment	(0.52)	(0.63)	(0.34)	(0.31)
	[0.25]	[0.25]	[0.39]	[0.26]
Total Employment	(1.21)	(1.23)	(0.73)	(0.66)
	[0.12]	[0.27]	[0.39]	[0.25]

Notes: Year at the top of the column is the base year. The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Appendix Table 1.J. The Corrected Standard Errors and Rubin Missingness Ratios for Summary Statistics in Appendix Table 1.A, Part 2

	1995		2000	
	23 Construction	**62 Health Care & Social Asst**	**23 Construction**	**62 Health Care & Social Asst**
0 to 5% Minority in BY:				
Growth in:				
Total Employment	(2.83)	(2.35)	(1.80)	(1.66)
	[0.21]	[0.37]	[0.42]	[0.40]
White Employment	(2.61)	(2.21)	(1.66)	(1.54)
	[0.21]	[0.36]	[0.41]	[0.39]
5 to 20% Minority in BY:				
Growth in:				
Total Employment	(2.00)	(1.72)	(1.12)	(1.15)
	[0.34]	[0.18]	[0.40]	[0.47]
White Employment	(1.58)	(1.51)	(0.90)	(0.98)
	[0.32]	[0.22]	[0.40]	[0.45]
20 to 50% Minority in BY:				
Growth in:				
Total Employment	(2.50)	(1.72)	(1.16)	(1.23)
	[0.16]	[0.35]	[0.26]	[0.25]
White Employment	(1.54)	(1.07)	(0.68)	(0.79)
	[0.16]	[0.33]	[0.18]	[0.26]
50 to 80% Minority in BY:				
Growth in:				
Total Employment	(3.84)	(5.13)	(2.49)	(2.00)
	[0.34]	[0.35]	[0.38]	[0.27]
White Employment	(1.58)	(2.16)	(1.00)	(0.87)
	[0.39]	[0.32]	[0.42]	[0.35]
80 to 100% Minority in BY:				
Growth in:				
Total Employment	(4.96)	(3.43)	(4.20)	(2.96)
	[0.13]	[0.20]	[0.28]	[0.44]
White Employment	(1.80)	(0.77)	(0.98)	(0.66)
	[0.35]	[0.28]	[0.21]	[0.57]

Notes: "BY" stands for base year. Year at the top of the column is the base year. The corrected standard errors and Rubin missingness ratios are computed following Appendix 1.7.D. The corrected standard errors are presented in parentheses. The Rubin missingness ratios are presented in brackets.

Chapter 2

Linking the Firms and Establishments in the Longitudinal Employer-Household Dynamics Infrastructure Data to the Census Business Register

2.1 Introduction

The crosswalk between the Longitudinal Employer-Household Dynamics (LEHD) infrastructure file system and the Census Business Register (BR) is authorized as part of the LEHD Infrastructure Project. This document describes the LEHD - BR crosswalk and its component inputs: the Business Register, Longitudinal Business Database (LBD), and the LEHD Infrastructure File system. [1] The output files include the LEHD - BR crosswalk at both the establishment and employer (business) levels. These output files facilitate linking a wide

[1] If you are reading this document on Census secure computers, then you should be accessing the Census Confidential version, which includes some information that cannot be included in the public version.

range of contextual variables relating to characteristics of the current and prior employers and co-workers of current employees.

The purpose of this technical documentation is to describe the linking process and to examine the properties and the qualities of the crosswalk files. Section 2.2 provides the general overview, update frequency, acquisition process, and the naming convention of the LEHD - BR crosswalk files. Section 2.3 provides a detailed introduction to the two primary input files used to create this crosswalk: the LEHD - Employer Characteristics File (ECF) system and the BR. The input file cleaning and preparation is also discussed here. Finally, Section 2.4 describes and assesses the algorithm used to perform the linking. False match and non-match rates for various populations are defined and estimated in order to examine the properties and quality of the LEHD - BR crosswalk output files.

2.2 Overview of the LEHD - BR crosswalk

2.2.1 Definition

The LEHD - BR crosswalk provides linking information at both the establishment level and employer levels. The LEHD employer-level data and employee (co-worker) characteristics data can be linked to the Census Bureau establishment and firm level micro-data via this crosswalk. In particular, unlike the LEHD Business Register crosswalk produced by Chiang et al. (2005), where the most detailed exact crosswalk is at the Employer Identification Number (EIN), State, 4-digit industry, and county level, the LEHD - BR crosswalk documented here provides a many-to-many match that links all of the establishments in the ECF on the LEHD side with all of their associated establishments on the BR side via the FAS_EIN.

FAS_EIN is the Employer Identification Number developed for LEHD Firm Age and Size (FAS) project, which added firm age and size to the Quarterly Workforce Indicators using national definitions of the employing firm. It is the firm identifier recorded in the LEHD 2011 production snapshot, the primary research version of the LEHD infrastructure file system.

The FAS_EIN was produced to allow the linkage of the LBD - Business Dynamics Statistics (BDS) micro-data to the LEHD Infrastructure File System. This linkage facilitated the creation of firm age and firm size variables in the LEHD ECF (Haltiwanger et al., 2014). Unfortunately, when that linkage was designed, it did not include vintage dating for the LBD input files that were used–making it very difficult to reverse the linkage and recover the correct LBD identifiers from the currently supported LBD vintages. [2] The EIN serves as the basis for the FAS_EIN. By using the FAS_EIN to create the LEHD - BR crosswalk, we attempt to replicate the link used in the production of the firm age and firm size variables in the LEHD infrastructure file system. The main difference here is that we pass directly to the Business Register, rather than passing through the LBD/BDS. The advantage of doing so is that we can avoid the vintage issues with the current LBD/BDS linkage, where the provenance of the LBD identifiers used for the original firm age and firm size variable integration cannot be verified.

2.2.2 Update Frequency

The Census Bureau's Employer Business Register is updated annually. The Census Bureau's Longitudinal Employer-Household Dynamics Infrastructure File System is updated on a quarterly basis. The LEHD - BR crosswalk is updated annually.

2.2.3 Acquisition Process

The creation of the LEHD - BR crosswalk files requires the presence of all BR files and the ECF from the LEHD Infrastructure File System.

2.2.4 Naming Conventions

Four output files are produced:

[2]This issue has now been addressed in the production of new LBD vintages, which get a proper vintage label that can be used to verify linkage provenance.

- lehd_br_susbm_xwalk_final_pre02.sas7bdat: the employer-level LEHD - BR crosswalk file by year for 1990 - 2001, [3] using single-unit firms (i.e., establishments) and submasters [4] from the BR Single Unit (SU) file only.

- lehd_br_susbm_xwalk_final_post02.sas7bdat: the employer-level LEHD - BR crosswalk file by year for 2002 - current, [5] using single-unit firms (i.e., establishments) and submasters from the BR Single Unit (SU) file only.

- lehd_br_sumu_xwalk_final_pre02.sas7bdat: the establishment-level LEHD - BR crosswalk by year for 1990 - 2001, using single-unit firms (i.e., establishments) from the BR Single Unit (SU) file and establishments that belong to multi-unit firms or submasters from the BR Multi-Unit (MU) file.

- lehd_br_sumu_xwalk_final_post02.sas7bdat: the establishment-level LEHD - BR crosswalk by year for 2002 - current, using single-unit firms (i.e., establishments) from the BR Single Unit (SU) file and establishments that belong to multi-unit firms or submasters from the BR Multi-Unit (MU) file.

2.3 Description of the Input Files & Initial Preparation

In this section, the two main input files: the LEHD - ECF file and the BR file are introduced. Additionally, the initial input file cleaning and preparation, mostly for the BR data, are also discussed in detail here.

[3] In 2002, the BR underwent a substantial redesign, changing to an entirely different identification numbering system, which is discussed in detail in Section 2.3.2. Therefore, it is necessary to create the LEHD - BR crosswalk files for 1990 - 2001 and 2002 - current separately.

[4] See section 2.3.2 for detailed description.

[5] The current version of the LEHD - BR Crosswalk is available for 2002 - 2012. Because in the LEHD 2011 production snapshot, 2012 quarter one is the last available quarter with data inputs, the current LEHD - BR crosswalk can only provide linkages between the 2012 BR entities and the LEHD entities in 2012 quarter one.

2.3.1 LEHD - Employer Characteristics File (ECF)

The first primary input file underlying the creation of the LEHD - BR crosswalk is the Employer Characteristics File (ECF) of the LEHD Infrastructure File System. The LEHD Infrastructure File System is a job-based frame and associated linked files designed to represent the universe of individual – employer pairs covered by state Unemployment Insurance (UI) system reporting requirement (with federal employees added in 2012). The underlying data of the LEHD Infrastructure File system are wage records extracted from the UI administrative files of each partner state in the Local Employment Dynamics (LED) cooperative federal – state partnership (Abowd et al., 2009). Besides the UI wage records, LED partner states also deliver an extract of the file reported to the Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW, formerly known as ES-202). These data are received by LEHD on a quarterly basis, with historical time series extending back to the early 1990s for many states. In the 2011 snapshot, LEHD Infrastructure File System has data from all 50 states in the U.S. A more comprehensive overview and description of the LEHD Infrastructure File System can be found in Abowd et al. (2009).

In the LEHD system, an employer is identified primarily by its state UI account number, which is recoded to SEIN – State Employer Identification Number. The SEIN is specific to a state. In LEHD, a single legal employer might have multiple SEINs, and, regardless of its operations in other states, a legal employer has a different UI account in each state in which it has statutory employees. [6] An establishment corresponds to a firm's reporting unit in ES-202. It is usually the place where the employees actually perform their work. Establishments in the LEHD Infrastructure File System are identified by SEIN concatenated with SEINUNIT. Most employers have one establishment. These firms are usually called single-units. However, most employment is with employers who have multiple establishments.

[6] That is to say, for example, a Target in New York and a Target in New Jersey are always considered different LEHD employers, but a Target in Ithaca, New York, and a Target in Binghamton, New York, may be considered be part of the same SEIN if they have the same owner and, therefore, the same state UI account.

These multi-establishment employers are usually called multi-units.

The ECF of the LEHD Infrastructure File System is itself a collection of related files. In the 2011 snapshot, the ECF consolidates most employer and establishment-level information into three files: the employer SEIN-level file, the establishment SEINUNIT-level file, and the newly added firm age and firm size file. The ECF establishment SEINUNIT-level file contains one record for every year-quarter in which the SEIN-SEINUNIT pair is present in either the ES-202 or the UI wage records. The ECF employer SEIN-level file contains one record for every year-quarter in which the SEIN is present in either the ES-202 or the UI wage records. It is built up from the ECF establishment SEINUNIT-level file. Therefore, the ECF employer SEIN-level file provides no additional information but is an easier and more efficient way to access SEIN-level summary data. The newly-added firm age and firm size file contains one record for every SEIN-SEINUNIT-year-quarter with national employer-level data and identifiers imported from the LBD/BDS micro-data.

A number of input files are used to build the SEIN-level and the SEINUNIT-level ECF files from the component the QCEW/ES-202 and the UI wage record data. In particular, the QCEW/ES-202 collects employment, payroll, economic activity, and physical location information from employers covered by state UI programs and from employers subject to the reporting requirements of the ES-202 system. The universe for these data is a reporting unit, which is the BLS's definition of an establishment. Over the years, the information contained in the SEIN-level and SEINUNIT-level files has increased substantially. Common to all years are information on the employer's identity (the SEIN), the reporting unit's identity (SEINUNIT), ownership information, employment on the twelfth of each month covered by the quarter, total wages paid over the course of the quarter, and information pertaining to industry classifications, including both the Standard Industrial Classification (SIC) and North American Industry Classification System (NAICS). Federal identifiers, (e.g., the Employer Identification Number (EIN)), and geographic identifiers are also included.

The creation of the firm age and firm size file relied primarily on data from the LBD and

BDS micro-data and the National Employer Characteristics File (NECF) augmented with the LEHD Successor-Predecessor (SPF) state files (Haltiwanger et al., 2014). The LBD/BDS contains annual longitudinal establishment information based on national firm-level entities for nearly the entire non-farm private economy and a small portion of the public sector from 1976 through the present. The NECF was generated by interleaving all the state ECFs. The SPFs identify spurious changes of the firm identifiers due to events such as mergers and acquisitions in the LEHD. The variables found in the ECF firm age and firm size file include LBD firm alpha, identifying the enterprise, [7] national firm size, [8] national firm (alpha) age, [9] and the critical identifier that permits the creation of the LEHD - BR crosswalk: FAS_EIN.

A Brief Discussion on the FAS_EIN

The LEHD - BR crosswalk relies on the identifier called FAS_EIN. In particular, to facilitate the creation of firm age and firm size variables, the LBD/BDS micro-data were enhanced by integrating the EIN from the Business Register. The result is the EIN input to the FAS_EIN variable: the FAS_EIN is the EIN reported in the BR LBD/BDS prepended with five zeroes that are subsequently edited by the LEHD staff to reflect discrepancies between the LBD/BDS EIN and the one reported on the QCEW (Haltiwanger et al., 2014).

The corresponding enhancement to the LEHD infrastructure files is not straightforward. The addition of FAS_EIN in the LEHD - ECF required correcting missing, partially missing, or invalid EINs as reported in the QCEW files delivered via the LED partnership (Haltiwanger et al., 2014). Once the EINs are edited, if an SEIN has a valid EIN, the FAS_EIN is created by pre-appending five zeros to the EIN. If an SEIN record contains no EIN in the ECF, the FAS_EIN is generated by appending a two-digit number to the SEIN. A more comprehensive overview and description of the LEHD firm age and firm size file and the

[7] An enterprise is the parent firm that controls more than 50% interest in the related establishments.

[8] This variable measures the best initial firm size as of March 12^{th} of the previous year, or current size if the firm was born in current year.

[9] This LEHD firm age variable was created as an edited version of the LBD variable such that there can be no abrupt changes in firm age at the alpha level due to activities such as merger, acquisition, or establishment entry and exit from the alpha firm(Haltiwanger et al., 2014).

creation of FAS_EIN can be found in Haltiwanger et al. (2014). Table 2.1 provides the percentage of FAS_EIN identifiers that are valid EINs [10] by year from 1990 to 2012. According to Table 2.1, approximately 95% or more of all FAS_EINs are valid in the LEHD - ECF each year, which suggests high utility from using of this identifier to link to EIN indexed data like the BR. [11]

The LEHD - ECF File Preparation

To create the LEHD - BR crosswalk, all three files in the LEHD - ECF file system are used: the SEIN-level file provides information on SEIN-level payroll, SEIN-level employment, number of units owned by a SEIN, and cleaned SEIN-level 2007 NAICS industry code; the SEINUNIT-level file provides detailed SEINUNIT-level information on geography; and finally, the firm age and firm size file provides FAS_EIN and the LBD/BDS firm identifier (alpha).

Due to the different data-reporting frequencies, the quarterly LEHD - ECF is annualized. Specifically, regarding the payroll variable, the annual payroll is simply the sum of all quarterly payrolls of a SEIN in a year. The quarterly employment variable is annualized by retaining the SEIN-level employment that is closest to March 12^{th}. For a SEIN - year: if the first available quarter is quarter 1, the annualized employment variable is set to equal to the month-three employment of quarter 1 of that year. However, if the first available quarter is quarter 2, 3, or 4 for the SEIN - year pair, the employment is set to equal to the month-one employment of the first available quarter of that year.

2.3.2 Business Register (BR)

The Business Register (BR), which was formerly known as the Standard Statistical Establishment List file (SSEL), is a central repository of legal entities (generally businesses) that

[10] A valid FAS_EIN, i.e., one that will link to other data indexed by the federal EIN, is the one created by pre-pending five zeros to a valid EIN.

[11] The results presented in Table 2.1 are very similar to those presented in Table A1 in Haltiwanger et al. (2014).

operate within the U.S. and its island territories as identified by the Master File systems of the U.S. Internal Revenue Service (IRS) (Census Bureau, 2013b; Jarmin and Miranda, 2002; Salyers, 2004). The primary function of the BR maintained by the U.S. Census Bureau is to develop the frames for economic censuses and surveys (Jarmin and Miranda, 2002). Additionally, the BR also serves as the central repository of administrative records (primarily federal tax data) and the main source of basic employment and payroll measures summarized by industry and geographic area in the annual County Business Patterns (CBP) and ZIP Business Patterns statistical series (Salyers, 2004).

Input Data Source and Suppliers

Administrative records are the foundation of the BR. the BR's principal administrative record supplier is the Internal Revenue Service (IRS). Specifically, the Business Master File Entity/Directory (BMF), the Employer's Quarterly Federal Tax Return – IRS Form 941 series (agricultural employers file the Employer's Annual Tax Return for Agricultural Employees – IRS Form 943 series), and the annual business income tax returns from the IRS are of primary importance. The BMF extracts provide the BR with EINs, legal and trade names, mailing and physical location addresses, principal business activity (industrial) classification codes, processing indicators, etc; the payroll tax returns data from the Employer's first-quarterly Federal Tax Return provides total employment for the pay period including March 12^{th} and total payroll for the quarter; the annual business income tax returns data provide basic measures of business receipts/revenue, assets, and a principal business activity (industrial) classification codes.

Besides the administrative records from the IRS, the Social Security Administration (SSA) and the Bureau of Labor Statistics (BLS) are also important record suppliers. New business and organizational taxpayers (i.e., births) file an Application for Employer Identification Number, Form SS-4, with the IRS. The Census Bureau receives data obtained via Form SS-4 from the SSA processing. Form SS-4 content supplied to the Census Bureau in-

cludes EIN, industry (NAICS) codes , geographic information, estimated employment, and other classification/status indicators. Finally, each quarter, the Census Bureau prepares a file of EINs that identifies unclassified single units and partially classified manufacturing single units from the BR. The BLS refers each of these EINs to the Business Establishment List (BEL) maintained by the BLS, which is QCEW-based, and returns the corresponding industry (NAICS) codes, whenever possible.

Statistical Units in the BR

The BR identifies four basic types of statistical units: establishment, EIN entity, enterprise, and alternative reporting unit. In the BR, an establishment is an economic unit, usually a single physical location, where business is conducted, or where services/industrial operations are performed. An EIN entity is an administrative unit that the IRS has assigned a unique identifier, i.e., the EIN, for tax reporting purposes. For single-unit establishments, the EIN entity and the establishment are the same. This is not the case, however, for establishments of multi-unit firms. A multi-unit firm usually has at least one EIN. Each EIN that a multi-unit firm reports under is flagged as a "submaster" (Jarmin and Miranda, 2002). An enterprise, which is also referred to as a parent company, is an economic unit comprising one or more establishment under common ownership or control. In the BR files, the enterprise is identified via a 6-digit number called "Alpha". The concepts of alternative reporting units are established by the Census Bureau specifically for data collection purposes for industries that cannot report establishment data. These units typically represent a part of the company made up of all activity within a given industry and geographic area.

The Organizational Structure of the BR Files and the BR Redesign

There is great variation in terms of the complexity of business organizations. Figure 2.1 shows an example of the complex organizational relationships among establishments, EINs, and alternative reporting units within an enterprise. This figure is borrowed from Salyers

(2004). For this particular reason, the BR files are organized based on the level of business organizational complexity. For each year, the BR has two files – SU and MU. [12] SU stands for "Single Unit" and MU stands for "Multi-Unit". The BR - SU file contains single establishment enterprise records. For a single establishment enterprise, the establishment, EIN entity, and the enterprise are the same. The SU file also contains the submaster records. These submasters are EIN-level records, *not* establishment-level records. The BR - MU file contains multi-unit establishment records. These establishments are owned by multi-unit enterprises or owned by submasters. For a multi-unit enterprise, the establishment and the legal entity are not coincident.

For illustration purposes, Figure 2.2 shows three types of simple organizational structures of companies: a single-unit establishment (Panel A), a multi-unit company with no submasters (Panel B), and a multi-unit company with submasters (Panel C). Using Figure 2.2 as an example, the data record for the single-unit establishment (Panel A) is stored in the SU file; all the establishment-level data records associated with the multi-unit company with no submasters (Panel B) are stored in the MU file; the data records of the submasters in Panel C are stored in the SU file, and the establishment-level data records associated with these submasters are stored in the MU file. In both the SU and the MU files, there is a 6-digit number – Alpha to identify the enterprise/parent company to which these establishments and submasters belong.

Before 2002, the old SSEL identified entities by EINs and Census File Number (CFN). A CFN is an unique identification number assigned to each establishment by the Census Bureau. The CFN is constructed differently depending on whether an entity is an SU establishment or MU establishment. For a single-unit company, its CFN is a 10-digit number consisting of a leading zero prepended to its 9-digit EIN. As a comparison, for a multi-unit company, its CFN is a 10-digit number with the first 6 digits being the Alpha identifier and

[12] This is true except for 1988. The 1988 MU file is lost forever, and is replaced by the Census Bureau's County Business Patterns (CBP) file. The CBP is an annual series that provides sub-national economic data by industry. This series includes the number of establishments, employment during the week of March 12, first quarter payroll, and annual payroll.

the last four digits being a location number. One issue with the CFN is that it lacks longitudinal consistency. The CFN is usually very consistent within a tax filing year across surveys, however, it is much less so across time (Salyers, 2004). Additionally, the old identification number system does not permit the organizational structures of firms to be fully and correctly retrieved. Using Figure 2.1 as an example again, under the old identification number system, we can identify all the establishments associated with an enterprise via the Alpha identifiers. However, we could not accurately link the establishments to their associated submasters.

In order to improve the support for business surveys and to strength its effectiveness in providing comprehensive and accurate coverage of the universe of business, in 2002, the BR underwent an extensive redesign. The BR was redesigned to allow the retention of an increased number of data elements that were previously lost. It was also redesigned to provide sufficient flexibility to fully accommodate the complex business organizations involving the four statistical units described above (Salyers, 2004).

The primary change in the BR redesign is the creation of a set of new identification numbers for register entities. Specifically, the CFN was discontinued and replaced by a 10-digit serial number that has no embedded meaning for each register entity. This new identification number is referred to as the "estab. identifier" in this paper due to the confidentiality of the variable name itself. Similar to CFN, this estab. identifier is supposed to be unique in the annual BR files. One major advantage of the new identifier is that it allows an establishment to maintain the same identification number irrespective of company organizational changes or changes in its tax filings. In addition to the estab. identifier, two more identification numbers are also developed to facilitate full identification of an establishment's parent firm and submaster. These three identifiers together permit the creation of a centralized "link" table that allows all register entities to be related to each other accurately. Refer to Figure 2.1 again, under the new identification number system after the 2002 BR redesign, we now can accurately link establishments to their associated submasters and enterprises.

The BR File Preparation

To create the LEHD - BR crosswalk, prior to the 2002 BR redesign, the SU and MU files provide four identifiers: CFN, EIN, Alpha, and permanent plant number (PPN). [13] After the 2002 BR redesign, the SU and MU files provide five identifiers: the three identifiers that identify an establishment, its associated enterprise, and its associated submaster, EIN, and Alpha. Besides these identifiers, the SU file provides administrative records on current year quarterly payroll (IRS Form 941), current year annual payroll for Agricultural Employees (IRS Form 943), current year employment, activity code, processing division code, and geographic information including state, county, and place FIPS code; and the MU file provides data on reported annual payroll, reported employment, activity code, processing division code, and geographic information including state, county, and place FIPS code.

Though the CFN and the estab. identifier are supposed to be unique and only one data record for every establishment is supposed to exist each year, there are duplicates in the BR files. Before creating the LEHD - BR crosswalk, all the BR SU and MU files were cleaned. Prior to the 2002 BR redesign, CFN duplication only exists in the 1993 SU and MU files. In Table 2.2, I summarize the unique CFNs that have duplicates and their frequencies. In the 1993 BR SU file, 56 CFNs have a frequency of 2, which account for 112 observations. In addition, 1 CFN has a frequency of 4, which accounts for 4 observations. In the 1993 BR MU file, 1 CFN has a frequency of 3, which accounts for 3 observations. The CFN with the frequency of 4 in the SU file and the CFN with the frequency of 3 in the MU file turns out to be the code for invalid or missing CFN. These observations were deleted. As far as the 56 CFNs with the frequency of 2 were concerned, it turns out that for each unique CFN, the duplicated observation has less information in the form of missing geographic code, payroll, and employment, etc. The solution for these CFNs is that only the observation with more information is retained. To identify the CFNs with the frequency of 2 and retrieve the

[13] PPN is also an establishment identification number assigned by the Census Bureau. However, unlike CFN, it is a longitudinal ID variable.

deleted observations from the original BR files, one can use the variable CFN_FREQ in the LEHD - BR crosswalk, which equals 2 for the establishments with duplicate CFNs.

After the 2002 BR redesign, similar duplicates also exist but only in the BR SU files. Table 2.3 shows the frequencies, the counts of the non-unique estab. identifiers and the total number of observations accounted for. These numbers are provided for single-unit establishments and submasters separately. As shown in Table 2.3, all non-unique estab. identifiers have a frequency of 2. A closer examination reveals that, for each non-unique estab. identifier, it is always the case that one observation is associated with an active establishment and the other observation is associated with an inactive establishment, defined by the BR establishment-level activity code. To resolve the duplication issue, for each non-unique estab. identifier, the observation associated with the active establishment is kept and the one associated with the inactive establishment is deleted. In order to identify these estab. identifiers with the frequency of 2 and retrieve the deleted observations from the original BR files, one can use the variable Estab_Identifier_FREQ in the LEHD - BR crosswalk, which equals 2 for the establishments with duplicate estab. identifiers.

Because the main functionality of the BR is to serve as the repository of legal entities (generally businesses) in the U.S., many data records in the BR files are there for processing reasons and should be excluded from statistical studies because they are considered inactive establishments. Active establishments are defined as those entities with positive employment and/or payroll within a year. However, given the difference in the update frequencies of the BR files and the LEHD - ECF files, as well as the different universes of establishments covered, to maximize the match validity, all BR entities are used for the creation of the LEHD - BR crosswalk, regardless of activity status. The activity status is flagged by the variable "active" in the crosswalk. This activity status variable is created using payroll, processing division code, and BR's original activity code.

2.4 The Creation of LEHD - BR Crosswalk

2.4.1 File Structure and Contents

In this section, the entity identifiers and the unit of observation for statistical analyses are discussed.

Identifiers

Two types of identifying variables in the construction of the LEHD - BR crosswalk files (both the employer-level and the establishment-level) are used. These are: business identifier and geographic information. The business identifier used is the FAS_EIN which is an enhanced EIN created for the purpose of incorporating firm age and firm size variables into the LEHD Infrastructure File System. [14] The key geographic information used to create the crosswalk is the two-digit state FIPS code.

Unit of Observation

Except for Minnesota, where wage records correspond to the report of an individual's UI-covered earnings by a firm's reporting unit, i.e., the SEINUNIT, most wage records in the LEHD Infrastructure File System correspond to the report of an individual's UI-covered earnings by an employing entity, i.e., the SEIN. As a result, the recommended level of analysis for most wage determination studies using the LEHD Infrastructure File System is the SEIN level. The unit of observation in the LEHD - BR crosswalk is a unique CFN/estab. identifier - FAS_EIN - State - SEIN record.

2.4.2 Algorithm

In this section, the algorithm used to create the LEHD - BR crosswalk is discussed in detail. Each step corresponds to a section of the SAS programs/bash scripts written to create the

[14] Details on how the FAS_EIN is created can be read in section 2.3.1 and Haltiwanger et al. (2014).

LEHD - BR crosswalk. The programs can be found on the Census Research Data Center server at:
//nsf-app1/rdcprojects/co/co00538/programs/zheng008/CenHRS/LEHD_BR_xcrosswalk/.
The LEHD-BR Crosswalk files can be found on the Census Research Data Center server at://nsf-app1/mixedtmp/co00538/CW_LEHD_BR/data.
The confidential version of this documentation can be found on the Census Research Data Center server at://nsf-app1/mixedtmp/co00538/CW_LEHD_BR/doc.

Step 1: Variables from the LEHD - ECF files are collected. In particular, from the LEHD - ECF firm age and firm size files, FAS_EIN, fas_ein_flag, fas_firm_id, quarter, sein, seinunit, and year are collected; from the LEHD - ECF SEIN-level files, SEIN, year, quarter, sein_best_emp1, sein_best_emp2, sein_best_emp3 (these three variables measure SEIN monthly employment), sein_best_wages, NUM_ESTABS, MODE_ES_NAICS_FNL2007_EMP, multi_unit, MODE_ES_COUNTY, MODE_ES_COUNTY_EMP, MODE_LEG_COUNTY, MODE_LEG_COUNTY_EMP are collected.

Step 2: Variables from the BR files are collected. Besides the necessary identifiers discussed in previous sections, variables from administrative records on current year quarterly payroll (IRS Form 941), current year annual payroll for Agricultural Employees (IRS Form 943), current year employment, activity code, processing division code, and geographic information including state, county, and place FIPS code are collected from the annual BR SU file; and the MU files provide data on reported annual payroll, reported employment, activity code, processing division code, and geographic information including state, county, and place FIPS code.

Step 3: Because in the original BR files, the same variable may have different length and format across years, this step is to reformat the variables collected from the BR files

such that the same variables have the same length and format to facilitate concatenation in the following steps. In addition, each year's SU file is split into two files: one containing only the single-unit establishments and one containing only the submasters. The variables active and active_flag are also created in this step. By completing this step, there are three BR input files for creating the LEHD - BR crosswalk each year: one contains only the single-unit establishments; one contains only the submasters; and the MU file.

Step 4: Diagnostics are run and issues with the CFN/estab. identifier duplication are discovered.

Step 5: The frequencies of CFNs and estab. identifiers are computed for each year's BR files and the frequency counts are stored.

Step 6: Diagnostics are run in order to characterize the non-unique establishment identifiers (CFN/estab. identifier). Details are summarized in Section 2.3.2 and Tables 2.2 and 2.3.

Step 7: Duplication of establishment identifiers in the BR files are edited. Details of the edit are discussed in Section 2.3.2. Tests are run and the results confirm that establishment identifiers are unique after Step 7.

Step 8: Tests are devised to examine whether we can accurately retrieve the firm structure involving enterprises, submasters and their associated establishments. It turns out that, as discussed in Section 2.3.2, only the BR files after the 2002 redesign permit such accurate linking.

Step 9 and Step 10: These two steps create the establishment-level BR input file and the employer-level BR input file for each year. To do so, first, tests are devised to

examine whether any CFNs/estab. identifiers exist in the file containing only the single-unit establishments, the file containing only the submasters, and the MU file simultaneously. It turns out that such CFNs/estab. identifiers do not exist. Then, the establishment-level BR input file is created using the file containing only the single-unit establishments and the MU file, and the employer-level BR input file is created using the file containing only the single-unit establishments and the file containing only the submasters. [15] For each year, the establishment-level BR input file contains all the establishments associated with all the enterprises (identified by the Alpha identifiers) in the BR. The employer-level BR input file contains all the single-unit establishments (enterprises with only one establishment) and all the submasters, which represent a subset of the establishments contained in the MU files.

Step 11: This step is to concatenate the annual BR files. Given the changes in the identification number system before and after the BR redesign, as well as the distinction between the establishment-level and the employer-level files, the annual BR files are concatenated into four final input files to create the LEHD - BR crosswalk: the establishment-level file before the BR redesign, the establishment-level file after the BR redesign, the employer-level file before the BR redesign, and the employer-level file after the BR redesign.

Step 12: This step is written to first convert the quarterly SEIN-SEINUNIT-level LEHD - ECF firm age and firm size file into an annual SEIN-level file. Then, the SEIN-level characteristics including payroll and employment and geographic information are supplemented using the LEHD - ECF SEIN - level file. The employment and geographic information that is closest to March 12^{th} is always used. Details on how the payroll and employment variables are retained are described in Section 2.3.1.

Step 13: This step is to create the simplest match at the FAS_EIN level by year. Match

[15] The employer-level BR input file is essentially the same as the original BR - SU file

rates are computed and presented in Section 2.4.3.

Step 14: To create the establishment-level and the employer-level LEHD - BR crosswalk at a finer level using both the FAS_EIN identifier and the geographic information, i.e., State, it is important to ensure that many-to-many record linking is not committed. To do so, in this step, the number of establishments or employers associated with a FAS_EIN - State pair is computed by year using the BR input files, [16] and the number of SEINs associated with a FAS_EIN - State pair is computed by year using the annualized SEIN-level LEHD - ECF file.

Step 15 and Step 16: Step 15 is to create the match at the FAS_EIN - State level by year. Match rates and FAS_EIN coverage rates are computed. The results are presented in Section 2.4.3. Using the frequency counts created in Step 14, in step 16 a flag is also created indicating which FAS_EIN - State pairs permit one-to-one or one-to-many record linking.

Step 17: Step 17 brings the FAS_EIN level match and the FAS_EIN - State level match and all the corresponding flags into one file. This is done separately for the employer-level files and the establishment-level files. The output files from this step serve as the basis in Step 18 for the creation of the final LEHD - BR crosswalk output files.

Step 18: Step 18 conducts the final crosswalk construction. It is done for 1990 - 2001 and 2002 - current, separately. Additionally, it is done at the establishment-level and the employer-level, separately.

Step 19: Recall, the establishment-level BR input files contain all the establishments associated with all the enterprises (identified by the Alpha identifiers) in the BR. The employer-level BR input files, on the other hand, contain all the single-unit establishments (enterprises

[16] This is done separately for the employer-level input files and the establishment-level input files.

with only one establishment) and all the submasters representing only a subset of establishments contained in the MU files. Therefore, the employer-level crosswalk is systematically missing the group of employers associated with the establishments contained in the BR - MU files that do not belong to any submasters. To assess the importance of these employers, weighted and unweighted percentages of establishments in the MU file that do not belong to any submasters are computed using all the establishments in the MU files as the universe. The weight used is the annual establishment-level reported employment. The results are presented in Section 2.4.3.

Step 20: False match rates and false non-match rates are calculated in Step 20. The methodology and the results are presented and discussed in Section 2.4.3.

2.4.3 LEHD - BR Match Rate

In this section, match rates at various levels are presented and discussed.

Match Rate at the FAS_EIN Level

The match rate at the FAS_EIN level by year is computed for both the establishment-level and the employer-level LEHD - BR crosswalks. It is done separately first using LEHD and then using the BR as the universe. The results are presented in Table 2.4. Columns (1) and (4) present the number of unique FAS_EINs that are matched between the LEHD and the BR. Columns (2) and (5) are computed using the valid FAS_EINs in LEHD as the universe. Columns (3) and (6) are computed using the (FAS_)EINs (that are associated with both active and inactive establishments or employers) in BR as the universe.

In the establishment-level LEHD - BR crosswalk, more than 90% of the valid FAS_EINs in LEHD can be matched to the BR each year (Column (2)). In comparison, a much lower percentages of (FAS_)EINs in the BR can be matched to the LEHD, especially in the earlier years when LEHD program had less participating states. For instance, in 1990, less than

30% of (FAS_)EINs in the BR can be matched. This percentage has gradually increased over the years, though it has always remained below 50% until the 2002 BR redesign. Column (3) shows a clear increase in 2002 in the percentage of (FAS_)EINs in the BR that can be matched to LEHD (from 39.7% in 2001 to 65.9% in 2002). Interestingly, after 2002, this percentage decreases annually from close to 66% in 2002 to lower than 40% in 2012.

One possible reason for such an unexpected trend could be that, contrary to what has been recommended, all the establishments are used in the creation of the LEHD - BR crosswalk regardless of its activity status. To reconcile the decline in FAS_EIN match rate after 2002, the FAS_EIN level match and the match rate computation are conducted again using only active-establishment-associated (FAS_)EINs in the BR. [17] The results are presented in Table 2.A. By comparing columns (3) in Table 2.4 and Table 2.A, when using only the active-establishment-associated (FAS_)EINs, the highest match rate between LEHD and BR is achieved in 2001, right before the 2002 BR redesign. In 2002, the match rate declines by more than 15%. And the declining trend in the match rate after 2002 is similar to that in Table 2.4. Therefore, it can be concluded that the unexpected decline in Table 2.4 is unlikely to be caused by using all (FAS_)EINs in the BR. Additionally, by comparing columns (2) in Table 2.4 and Table 2.A, we can see that using only the active-establishment-associated (FAS_)EINs usually renders lower match rates in LEHD. It verifies that by using all establishments from the BR, higher match validity is indeed achieved.

A close comparison between column (1) and (4), column (2) and (5), and column (3) and (6) in Table 2.4 indicates that the yearly match rate at the FAS_EIN level in the employer-level LEHD - BR crosswalk mirrors almost exactly the same pattern as what is described in the establishment-level LEHD - BR crosswalk.

[17] The activity status is defined by the variable "active".

Match Rate at the FAS_EIN - State Level

The match rate at the FAS_EIN - State level and the FAS_EIN coverage rate are computed for both the establishment-level and the employer-level LEHD - BR crosswalks, again using the LEHD and BR universes separately. The results for the establishment-level LEHD - BR crosswalk are presented in Table 2.5 (Columns (1) - (3)) and Table 2.6 (Columns (1) - (3)). The results for the employer-level LEHD - BR crosswalk are presented in Table 2.7 (Columns (1) - (3)) and Table 2.8 (Columns (1) - (3)). In addition, in these four tables, the percentage of FAS_EIN - State pairs that can be one-to-one or one-to-many linked between LEHD and BR and the corresponding FAS_EIN coverage rate are also presented (Columns (4) - (6) in Tables 2.5, 2.6, 2.7, 2.8).

In Table 2.5, Column (1) presents the number of unique FAS_EIN - State pairs that can be successfully matched between LEHD and BR in the establishment-level LEHD - BR crosswalk. Columns (2) and (3) compute the match rates of FAS_EIN - State pairs using the LEHD and the BR as the universe separately. For instance, in LEHD (Column (2)), the matched FAS_EIN - State can account for more than 85% of all FAS_EIN - State pairs in LEHD each year before 2002. In 2002, we observe a more-than-20-percent decrease in terms of match rate (from 85.5% in 2001 to 63.6% in 2002). Since 2002, the match rate has gradually increased to close to 78% in 2012. In BR ((column (3)), on the other hand, we observe a significant lower match rate, especially in the earlier years. It gradually increases to close to 40% with 49.1% being the highest match rate which is observed in 2004.

The FAS_EIN coverage rate, which measures the percentages of the unique FAS_EINs that can be accounted for by the FAS_EIN - State match in the establishment-level LEHD - BR crosswalk, is presented in Table 2.6. Specifically, Column (1) presents the count of FAS_EINs covered by the matched FAS_EIN - State pairs (i.e., Column (1) in Table 2.5). Columns (2) and (3) compute the FAS_EIN coverage rates using the FAS_EINs in LEHD and BR as the universe separately. Between 1990 and 2001, in LEHD, more than 90% of FAS_EINs are covered by the FAS_EIN - State match each year. Similar to the FAS_EIN -

State match rate, this coverage rate decreases substantially in 2002 and has remained lower since then. In the BR, usually between 30% and 48% of the (FAS_)EINs are covered by the FAS_EIN - State match each year. A close comparison between Columns (2) and (3) in Table 2.6 and those in Table 2.4 reveals that each year only a small fraction of FAS_EINs in LEHD and BR cannot be matched once geographic information (state) is used to further refine the match.

Not every matched FAS_EIN - State pair yields feasible record linkage between LEHD and BR. It depends on the number of entities associated with a particular FAS_EIN - State pair in both input files. Columns (4) - (6) in Table 2.5 repeat the exercise done for Columns (1) - (3), but for those FAS_EIN - State pairs that permit one-to-one or one-to-many record linking between LEHD and BR. Comparing Columns (4) - (6) to Columns (1) - (3) shows that the vast majority of FAS_EIN -State pairs that can be matched between LEHD and BR do permit one-to-one or one-to-many record linking. The FAS_EIN coverage rates of these FAS_EIN - State pairs in LEHD and BR are presented in Columns (4) - (6) in Table 2.6. As we can see, in terms of magnitude, they are quite similar to those listed in Columns (2) and (3) in the same table, respectively. The exercise is repeated for the employer-level LEHD - BR crosswalk. The results are presented in Table 2.7 and Table 2.8. Overall, both the trends and the magnitudes of the match rates and FAS_EIN coverage rates are very similar to the establishment-level LEHD - BR crosswalk.

Recall that in Step 19 of Section 2.4.2, it is stated that the employer-level crosswalk systematically misses the group of employers associated with those establishments in the BR - MU files that do not belong to any submasters. To assess the importance of these employers, weighted and unweighted percentages of establishments in the MU file that do not belong to any submasters are computed using all the establishments in the MU files as the universe. The weight used is the annual establishment-level reported employment. The results are presented in Table 2.9.

In 1990 - 2001, the unweighted percent of MU establishments that do not belong to

any submasters are almost always below 10%. The unweighted percentage gets relatively higher after 2002 and remains between 20 - 40% each year. Though the magnitude becomes alarmingly large, once it is weighted by the annual establishment-level reported employment, these establishments/employers become rather insignificant. Column (3) in Table 2.9 shows that the weighted percent of establishments in the BR - MU file that do not belong to any submasters always remains below 5%. The weighted percentage becomes negligible after the 2002 BR redesign and always remains below 0.5%. Therefore, although the employer-level crosswalk systematically misses this group of employers, it appears to be an insignificant portion of employment and therefore should not affect our ability to capture the vast majority of employers in this level of the crosswalk files.

False Match & False Non-Match Rate at the FAS_EIN - State Level

To further evaluate the quality of the FAS_EIN - State match, false match rate and false non-match rate at the FAS_EIN - State level are computed for 2002. 2002 is chosen as a reference year because: (1) it is an economic census year; (2) it is the year in which BR was redesigned; (3) it is the year in the middle of 1990 - 2012. Company name, county, physical and mailing address are used to facilitate the manual comparison and the identification of the false matches and false non-matches. The variables from the 2002 BR - SU and BR - MU files that contain this information include: company names, county code for physical address, mailing address city, physical address city, mailing address street, and physical address street.

The variables containing the similar information in the LEHD Infrastructure File System come from the quarterly SEIN-SEINUNIT-level ES-202 files. These variables are: name_legal (legal/corporate name), name_trade (trade/DBA name), name_worksite (worksite name, i.e., reporting unit description), county (county code), address_city (physical location address city), other_address_city (mailing/other address city), address_street1 (physical location address street1), address_street2 (physical location address street2), other_address_street1

(mailing/other address street1), other_address_street2 (mailing/other address street2).

In order to compute the false match rate, all the SEINs that can be linked to a BR establishment are stratified into three strata using the SEIN-level employment measure (variable sein_best_emp): small (less than 20 employees), medium (20 - 499 employees), and large (500+ employees). Then, 300 firms are randomly sampled from each stratum. This is done for the establishment-level and the employer-level crosswalk files separately.

To merge in the company name, county, and address information from the BR, the stratified random sample of 900 SEINs from the establishment-level crosswalk file are first merged back to the establishment-level LEHD - BR crosswalk to obtain all the establishment identifiers and its associated unique line number. Then, the BR company name, county, and address variables are merged in using the establishment identifiers and its associated unique line number. The same steps are repeated for the employer-level crosswalk file to bring in the BR company name and address information.

Because the LEHD ES-202 files are quarterly and the LEHD - BR crosswalk files are annual, similar annualization as described in Section 2.3.1 is conducted and the company name and address information of the quarter that is closest to March 12^{th} is retained. Additionally, because the LEHD ES-202 files are at the SEIN-SEINUNIT level and the LEHD - BR crosswalk files are at the SEIN level, the company name and address information of the establishments (i.e., SEINUNITs) that are located in the SEIN employment mode cleaned county (MODE_ES_COUNTY_EMP) is merged in.

To determine whether a linked SEIN - FAS_EIN - State - BR establishment is a false match, a fuzzy logic is applied. In short, for an SEIN - FAS_EIN - State - BR establishment record, as long as the company name and county information from the LEHD ES-202 file matches the company name and county information of any establishments from the BR file, it is considered to be a true match. Otherwise, it is considered to be a false match. Since the company name and address information of the establishments located in other counties (i.e. counties besides the SEIN employment mode cleaned county) are not used, for the purpose

of identifying the false match cases, the company name and county information is considered to be primary. The actual physical or mailing address is considered to be secondary. That is to say, for a linked SEIN - FAS_EIN - State - BR establishment, if the county and company names are matched, but the physical and mailing address from the BR and the LEHD ES-202 files do not match, it is still considered as a true match. If company name, county, and address information is all missing either in the LEHD ES-202 file or in the BR file, it is considered to be an unknown case. The false match rate for the establishment-level and the employer-level LEHD - BR crosswalk are presented in the upper panel of Table 2.10.

Overall, in the establishment-level crosswalk file, the random sample of 900 SEINs results in 8158 SEIN - FAS_EIN - State - BR establishment pairs. The lower bound of the false match rate is calculated using the identified false match cases alone. The upper bound of the false match rate is calculated using both the identified false match cases and the unknown cases. As seen in Table 2.10, the lower bound of the false match rate in the establishment-level crosswalk file never exceeds 1%. If all the unknown cases were false matches, the false match rate for the small firms is 13.3% and the false match rates for the medium size firms and the large firms are always below 10%. In comparison, the random sample of 900 SEINs from the employer-level crosswalk file results in 901 SEIN - FAS_EIN - State - BR establishment pairs. The false match rate is very similar to that of the establishment-level crosswalk in terms of magnitude.

In order to compute the false non-match rate, all the SEINs that cannot be linked to the BR are stratified into three strata using the SEIN-level employment measure (variable sein_best_emp): small (less than 20 employees), medium (20 - 499 employees), and large (500+ employees). Then, 300 firms are randomly sampled from each stratum. This is done for the establishment-level and the employer-level crosswalk files, separately. Then, the company name, county, and address information from the LEHD ES-202 files are merged in the same manner as done for the false match rate calculation. The false non-match rate for the establishment-level and the employer-level LEHD - BR crosswalk files are presented in

the lower panel of Table 2.10.

Overall, in the establishment-level crosswalk file, the random sample of 900 SEINs results in 2716 SEIN - SEINUNIT pairs. The similar fuzzy logic described above is also applied. That is to say, as long as the company name and county information matches the company name and county information of any establishments from the BR file, it is considered as a false non-match. Otherwise, it is considered to be a true non-match case. Again, the company name and county information instead of physical and mailing address is placed with higher weight. The lower bound of the false non-match rate is calculated using the identified false non-match cases alone. The upper bound of the false non-match rate is calculated using both the identified false non-match cases and the unknown cases. As seen in Table 2.10, the lower bound of the false non-match rate in the establishment-level crosswalk file is 44% for the small firms, 60.3% for the medium size firms and more than 73% for the large firms. The upper bound of the false non-match rates for these three groups of firms are 66.7%, 73.3%, and 81.3%, respectively. The random sample of 900 SEINs results in 3591 SEIN - SEINUNIT pairs in the employer-level crosswalk file. The magnitude of the false non-match rates is very similar as compared to that of the establishment-level crosswalk file.

Taking the false match rates and non-match rates together, it can be concluded that, if an SEIN - FAS_EIN - State - BR establishment match is formed in the LEHD - BR crosswalk, it is likely to be a very accurate match. However, for those SEINs in the LEHD that cannot be linked to any establishment in the BR, it is highly likely that it is a false non-match. The reason detected for such high false non-match rates is the high rate of missing two-digit state FIPS codes in the BR. Therefore, if the two-digit state FIPS code quality in the BR were to increase, an even higher match rate can be achieved.

2.4.4 How to Use the LEHD - BR crosswalk

Because the LEHD - BR crosswalk is created at two different levels – the establishment-level and the employer-level, it is up to researchers to choose the level of analysis they want to

conduct. Once it is determined, researchers may use the appropriate crosswalk file. To select the part of the crosswalk where successful linkages between BR establishments and SEINs in the LEHD are formed, set flag_fas_ein = 'M' and flag_fas_ein_st = 'M' and fas_ein_st_1toM_flag = '1'. To select all the data records from the BR, set flag_fas_ein != 'L' and flag_fas_ein_st != 'L'. To select all the data records from the LEHD - ECF, set flag_fas_ein != 'B' and flag_fas_ein_st != 'B'.

2.5 Appendix

2.5.A. List of Acronyms

BDS: Business Dynamics Survey

BEL: Business Establishment List

BLS: Bureau of Labor Statistics

BMF: Business Master File Entity/Directory

BR: Business Register

CBP: County Business Patterns

CFN: Census File Number

ECF: Employer Characteristics File

EIN: Employer Identification Number

FAS: Firm Age and Size

FAS_EIN: Employer Identification Number developed for FAS project

HRS: Health and Retirement Study

IRS: Internal Revenue Service

LBD: Longitudinal Business Dynamics

LEHD: Longitudinal Employer - Household Dynamics

MU: Multi-Unit

NAICS: North American Industry Classification System

NECF: National Employer Characteristics File

PPN: Permanent Plant Number

QCEW: Quarterly Census of Employment and Wages

SEIN: State Employer Identification Number (State-level UI account identifier)

SEINUNIT: SEIN Establishment Identifier

SIC: Standard Industrial Classification

SPF: Successor-Predecessor File

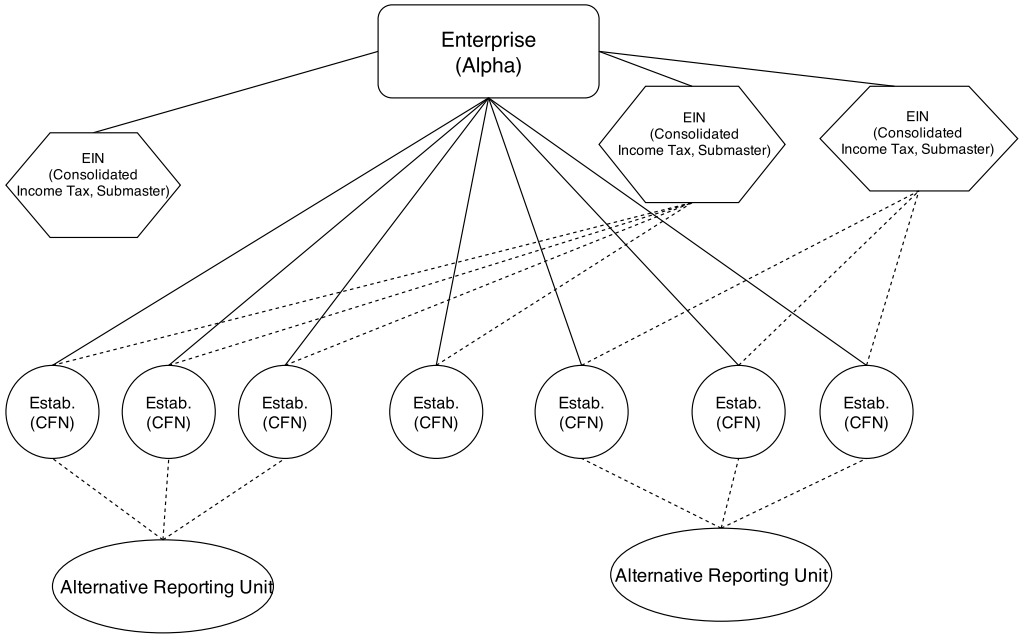
SSA: Social Security Administration

SSEL: Standard Statistical Establishment List

SU: Single-Unit

UI: Unemployment Insurance

Figure 2.1. Example of Statistical Unit Relations for a Small Multiple Establishment Company



Source: Salyers (2004).

Figure 2.2. Simple Organizational Structures of Firms in the Business Register

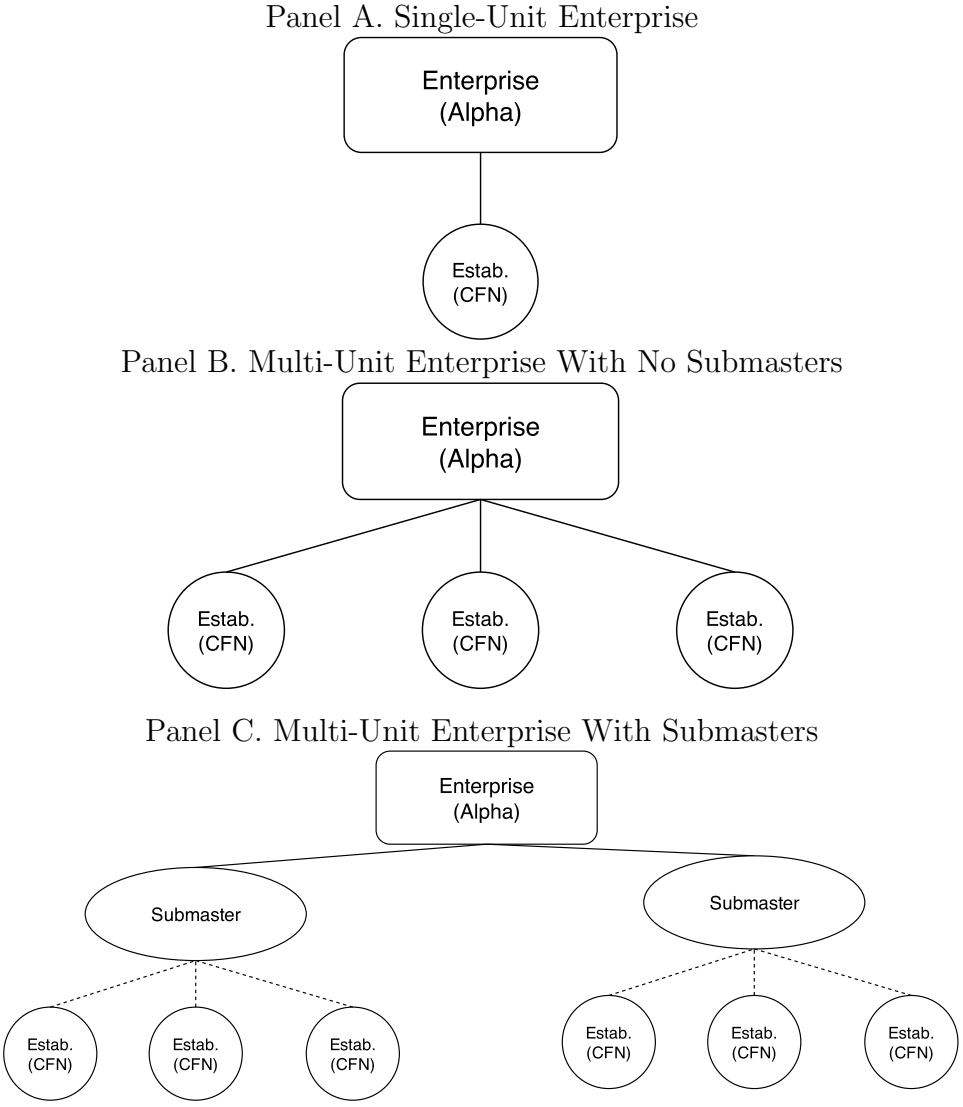


Table 2.1. Valid FAS_EIN By Year in LEHD - ECF

Year	Valid FAS_EIN Percent (1)	Invalid FAS_EIN Percent (2)	All FAS_EIN Percent (3)
1990	94.7	5.28	100.0
1991	95.9	4.15	100.0
1992	95.4	4.57	100.0
1993	95.7	4.31	100.0
1994	95.9	4.09	100.0
1995	96.4	3.65	100.0
1996	96.1	3.93	100.0
1997	96.4	3.65	100.0
1998	97.1	2.95	100.0
1999	97.4	2.57	100.0
2000	97.2	2.79	100.0
2001	97.7	2.31	100.0
2002	98.1	1.89	100.0
2003	98.4	1.62	100.0
2004	98.4	1.56	100.0
2005	98.6	1.43	100.0
2006	98.7	1.32	100.0
2007	98.7	1.28	100.0
2008	98.7	1.26	100.0
2009	98.8	1.17	100.0
2010	98.8	1.25	100.0
2011	98.0	2.02	100.0
2012	97.7	2.26	100.0

Notes: Data from the LEHD 2011 Snapshot are used. The percentages are calculated using the LEHD - ECF firm age and firm size file. A valid FAS_EIN is defined as the one created directly using EIN by pre-pending five zeros. All other FAS_EINs are defined as invalid FAS_EINs. The percentages do not sum to 100 due to rounding, which was done for disclosure avoidance purposes.

Table 2.2. Summary of the CFN Duplication in the 1993 Business Register SU & MU Files

File Type	CFN Frequency (CFN_FREQ)	Unique CFN	No. of Observation
SU	2	56	112
SU	4	1	4
MU	3	1	3

Notes: Data from the 1993 BR SU and MU files are used.

Table 2.3. Summary of the Establishment Identifier Duplication in the Business Register SU Files, 2002-2012

Year	Entity Type	Estab. Identifier Frequency (Estab_Identifier_FREQ)	Unique Estab. Identifier	No. of Observation
2002	Establishment	2	3684	7368
2002	Submaster	2	1	2
2003	Establishment	2	3768	7536
2003	Submaster	2	15	30
2004	Establishment	2	3828	7656
2004	Submaster	2	22	44
2005	Establishment	2	3860	7720
2005	Submaster	2	22	44
2006	Establishment	2	3844	7688
2006	Submaster	2	26	52
2007	Establishment	2	3788	7576
2007	Submaster	2	63	126
2008	Establishment	2	3780	7560
2008	Submaster	2	64	128
2009	Establishment	2	3779	7558
2009	Submaster	2	61	122
2010	Establishment	2	3714	7428
2010	Submaster	2	59	118
2011	Establishment	2	3700	7400
2011	Submaster	2	61	122
2012	Establishment	2	3662	7324
2012	Submaster	2	84	168

Notes: Data from the 2002 - 2012 BR SU files are used.

Table 2.4. Match Rate at the FAS_EIN Level, By Year

Year	Establishment-Level Match			Employer-Level Match		
	Universe: (Valid) FAS_EIN in					
		LEHD	BR		LEHD	BR
	Count	Percent	Percent	Count	Percent	Percent
	(1)	(2)	(3)	(4)	(5)	(6)
1990	2420000	94.9	27.6	2420000	94.9	27.7
1991	3350000	95.2	36.9	3350000	95.2	37.0
1992	3440000	95.1	35.7	3440000	95.1	35.8
1993	3540000	95.0	33.4	3540000	95.0	33.5
1994	3880000	94.3	36.1	3880000	94.3	36.2
1995	4310000	94.8	37.6	4310000	94.8	37.7
1996	4570000	93.8	44.6	4570000	93.8	44.7
1997	4650000	94.3	43.6	4650000	94.2	43.6
1998	5270000	94.4	45.0	5270000	94.4	45.1
1999	5530000	95.0	43.2	5530000	95.0	43.3
2000	5740000	95.2	41.7	5740000	95.2	41.8
2001	5860000	95.8	39.7	5860000	95.8	39.8
2002	5940000	93.4	65.9	5930000	93.3	66.5
2003	6120000	93.8	64.2	6110000	93.7	64.8
2004	6200000	93.9	61.6	6200000	93.9	62.1
2005	6330000	93.9	59.0	6330000	93.9	59.5
2006	6400000	93.9	56.6	6400000	93.9	57.0
2007	6460000	93.9	53.8	6460000	93.9	54.1
2008	6440000	93.6	51.4	6440000	93.6	51.7
2009	6300000	93.4	48.3	6300000	93.4	48.6
2010	6230000	93.0	46.0	6220000	93.0	46.3
2011	6170000	92.7	44.1	6170000	92.7	44.3
2012	5630000	93.3	38.5	5630000	93.3	38.7

Notes: Data from the 2011 LEHD Snapshot and 1990 - 2012 BR files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. The universe used to compute the match rate in Columns (2) and (5) is all valid FAS_EIN in LEHD by year. The universe used to compute the match rate in Columns (3) and (6) is all EIN in the BR files by year, after the duplicates are edited.

Table 2.5. Match Rate at the FAS_EIN - State Level in the Establishment-Level LEHD - BR Crosswalk, By Year

Year	Matched			One-to-One/Many Matched		
	Universe: FAS_EIN - State in			Universe: FAS_EIN - State in		
		LEHD	BR		LEHD	BR
	Count	Percent	Percent	Count	Percent	Percent
	(1)	(2)	(3)	(4)	(5)	(6)
1990	2390000	88.2	26.7	2390000	88.0	26.7
1991	3330000	89.1	36.1	3330000	88.9	36.0
1992	3440000	89.3	35.0	3440000	89.3	35.0
1993	3450000	86.9	32.0	3450000	86.8	31.9
1994	3890000	87.8	35.4	3880000	87.8	35.3
1995	4330000	87.8	37.0	4330000	87.7	37.0
1996	4580000	86.1	43.9	4580000	86.1	43.9
1997	4680000	86.6	43.0	4680000	86.5	42.9
1998	5320000	85.8	44.5	5320000	85.8	44.4
1999	5580000	85.8	42.7	5580000	85.7	42.7
2000	5790000	85.3	41.3	5790000	85.3	41.2
2001	5920000	85.5	39.3	5920000	85.5	39.3
2002	4580000	63.6	49.0	4580000	63.6	49.0
2003	4820000	65.2	48.8	4820000	65.1	48.7
2004	5130000	68.4	49.1	5130000	68.4	49.0
2005	5360000	70.0	48.2	5360000	69.9	48.2
2006	5560000	71.4	47.3	5560000	71.4	47.3
2007	5900000	74.7	47.2	5900000	74.7	47.2
2008	5880000	74.2	45.1	5870000	74.1	45.1
2009	5790000	74.4	42.8	5790000	74.3	42.8
2010	5740000	74.0	40.8	5740000	74.0	40.8
2011	5720000	73.7	39.3	5710000	73.7	39.3
2012	5430000	77.5	35.7	5420000	77.5	35.7

Notes: Data from the 2011 LEHD Snapshot and 1990 - 2012 BR files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. Column (4) is a subset of Column (1), however, due to rounding and the small number of FAS_EIN - State associated with multiple entities in both LEHD and BR, it is not obvious. The universe used to compute Columns (2) and (5) is valid FAS_EIN - State in LEHD by year. The universe used to compute Column (3) and (6) is EIN - State in the Establishment-level BR files by year, after the duplicates are edited.

Table 2.6. FAS_EIN Coverage Rate in the FAS_EIN - State Level Match Using Establishment-Level LEHD - BR Crosswalk, By Year

Year	Matched			One-to-One/Many Matched		
	Universe: FAS_EIN in			Universe: FAS_EIN in		
		LEHD	BR		LEHD	BR
	Count	Percent	Percent	Count	Percent	Percent
	(1)	(2)	(3)	(4)	(5)	(6)
1990	2340000	91.9	26.8	2340000	91.8	26.7
1991	3270000	92.9	36.0	3270000	92.8	36.0
1992	3370000	93.2	35.0	3370000	93.2	35.0
1993	3380000	90.7	31.9	3370000	90.7	31.9
1994	3800000	92.3	35.3	3800000	92.2	35.3
1995	4220000	92.9	36.9	4220000	92.9	36.9
1996	4480000	91.8	43.7	4470000	91.7	43.7
1997	4560000	92.5	42.7	4560000	92.5	42.7
1998	5180000	92.7	44.2	5180000	92.7	44.1
1999	5430000	93.1	42.4	5430000	93.1	42.4
2000	5630000	93.3	40.9	5630000	93.3	40.9
2001	5750000	93.8	38.9	5750000	93.8	38.9
2002	4390000	69.0	48.7	4390000	69.0	48.7
2003	4620000	70.9	48.6	4620000	70.9	48.5
2004	4930000	74.7	49.0	4930000	74.7	49.0
2005	5170000	76.6	48.1	5160000	76.6	48.1
2006	5360000	78.5	47.3	5350000	78.5	47.3
2007	5680000	82.5	47.2	5680000	82.5	47.2
2008	5660000	82.2	45.1	5650000	82.2	45.1
2009	5570000	82.6	42.7	5570000	82.6	42.7
2010	5520000	82.4	40.8	5520000	82.4	40.8
2011	5490000	82.5	39.2	5490000	82.5	39.2
2012	5210000	86.3	35.6	5210000	86.3	35.6

Notes: Data from the 2011 LEHD Snapshot and 1990 - 2012 BR files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. Columns (1) and (4) count the unique FAS_EIN in FAS_EIN - State pairs presented in Columns (1) and (4) in Table 2.5, respectively. The universe used to compute Columns (2) and (5) is valid FAS_EINs in LEHD by year. The universe used to compute Columns (3) and (6) is EINs in the Establishment-level BR files by year, after the duplicates are cleaned.

Table 2.7. Match Rate at the FAS_EIN - State Level in the Employer-Level LEHD - BR Crosswalk, By Year

Year	Matched			One-to-One/Many Matched		
	Universe: FAS_EIN - State in			Universe: FAS_EIN - State in		
		LEHD	BR		LEHD	BR
	Count	Percent	Percent	Count	Percent	Percent
	(1)	(2)	(3)	(4)	(5)	(6)
1990	2320000	85.6	26.6	2320000	85.6	26.6
1991	3250000	86.8	35.9	3250000	86.8	35.9
1992	3350000	86.8	34.8	3350000	86.8	34.8
1993	3340000	84.2	31.7	3340000	84.2	31.7
1994	3770000	85.3	35.2	3770000	85.3	35.2
1995	4200000	85.2	36.7	4200000	85.2	36.7
1996	4450000	83.8	43.5	4450000	83.8	43.5
1997	4540000	83.9	42.6	4540000	83.9	42.6
1998	5150000	83.2	44.0	5150000	83.2	44.0
1999	5400000	83.0	42.3	5400000	83.0	42.3
2000	5600000	82.6	40.8	5600000	82.6	40.8
2001	5720000	82.7	38.8	5720000	82.7	38.8
2002	4270000	59.2	47.8	4270000	59.2	47.8
2003	4510000	60.9	47.7	4510000	60.9	47.7
2004	4820000	64.3	48.3	4820000	64.3	48.3
2005	5060000	66.0	47.5	5060000	66.0	47.5
2006	5260000	67.6	46.8	5260000	67.6	46.8
2007	5590000	70.8	46.7	5590000	70.8	46.7
2008	5570000	70.2	44.6	5570000	70.2	44.6
2009	5490000	70.4	42.3	5490000	70.4	42.3
2010	5440000	70.1	40.3	5440000	70.1	40.3
2011	5480000	70.6	39.3	5480000	70.6	39.3
2012	5190000	74.2	35.6	5190000	74.2	35.6

Notes: Data from the 2011 LEHD Snapshot and 1990 - 2012 BR files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. Column (4) is a subset of Column (1), however, due to rounding and the small number of FAS_EIN - State records associated with multiple entities in both LEHD and BR, it is not obvious. The universe used to compute Columns (2) and (5) is valid FAS_EIN - State in LEHD by year. The universe used to compute Columns (3) and (6) is EIN - State in the Employer-level BR files by year, after the duplicates are cleaned.

Table 2.8. FAS_EIN Coverage Rate in the FAS_EIN - State Level Match Using Employer-Level LEHD - BR Crosswalk, By Year

Year	Matched			One-to-One/Many Matched		
	Universe: FAS_EIN in			Universe: FAS_EIN in		
		LEHD	BR		LEHD	BR
	Count	Percent	Percent	Count	Percent	Percent
	(1)	(2)	(3)	(4)	(5)	(6)
1990	2320000	91.1	26.6	2320000	91.1	26.6
1991	3250000	92.3	35.9	3250000	92.3	35.9
1992	3350000	92.5	34.8	3350000	92.5	34.8
1993	3340000	89.8	31.7	3340000	89.8	31.7
1994	3770000	91.6	35.2	3770000	91.6	35.2
1995	4200000	92.4	36.7	4200000	92.4	36.7
1996	4450000	91.3	43.5	4450000	91.3	43.5
1997	4540000	92.0	42.6	4540000	92.0	42.6
1998	5150000	92.3	44.0	5150000	92.3	44.0
1999	5400000	92.7	42.3	5400000	92.7	42.3
2000	5600000	93.0	40.8	5600000	93.0	40.8
2001	5720000	93.5	38.8	5720000	93.5	38.8
2002	4270000	67.1	47.8	4270000	67.1	47.8
2003	4510000	69.1	47.7	4510000	69.1	47.7
2004	4820000	73.0	48.3	4820000	73.0	48.3
2005	5060000	75.1	47.5	5060000	75.1	47.5
2006	5260000	77.1	46.8	5260000	77.1	46.8
2007	5590000	81.1	46.8	5590000	81.1	46.8
2008	5570000	80.9	44.7	5570000	80.9	44.7
2009	5490000	81.3	42.3	5490000	81.3	42.3
2010	5430000	81.2	40.4	5430000	81.2	40.4
2011	5480000	82.2	39.3	5480000	82.2	39.3
2012	5190000	86.0	35.6	5190000	86.0	35.6

Notes: Data from the 2011 LEHD Snapshot and 19905 - 2012 BR files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. Columns (1) and (4) count the unique FAS_EIN in FAS_EIN - State pairs presented in Columns (1) and (4) in Table 2.5, respectively. The universe used to compute Columns (2) and (5) is valid FAS_EINs in LEHD by year. The universe used to compute Columns (3) and (6) is EINs in the Employer-level BR files by year, after the duplicates are edited.

Table 2.9. Weighted and Unweighted Percentage of Establishments in the BR - MU File That Does Not Belong to a Submaster, By Year

Year	(Unweighted) Count (1)	(Unweighted) Percent (2)	(Weighted) Percent (3)
1990	216000	11.6	4.56
1991	66300	4.05	1.46
1992	121000	6.19	1.11
1993	153000	7.57	1.20
1994	198000	8.87	1.29
1995	251000	10.7	1.20
1996	77400	4.35	1.52
1997	162000	7.55	1.27
1998	235000	10.1	0.84
1999	319000	12.5	0.93
2000	400000	14.6	0.88
2001	607000	20.4	2.15
2002	1310000	40.9	0.45
2003	1310000	39.3	0.43
2004	1310000	37.8	0.33
2005	1310000	36.5	0.31
2006	1310000	34.9	0.28
2007	1300000	32.4	0.01
2008	1300000	31.2	0.15
2009	1300000	30.4	0.15
2010	1310000	29.5	0.15
2011	1310000	28.7	0.15
2012	1310000	27.4	0.13

Notes: Data from the 1990 - 2012 BR - MU files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. The universe used to computed Columns (2) and (3) is all establishments in the BR - MU files by year, after the duplicates are edited. Column (3) is weighted by establishment-level reported employment.

Table 2.10. False Match Rate and False Non-Match Rate in 2002

Level	Firm Size Group	Observation	Unique SEIN	False Match Rate Lower Bound	False Match Rate Upper Bound
Establishment-level	0 - 19	318	300	0.00 (0.006)	0.133 (0.017)
	20 - 499	713	300	0.0067 (0.006)	0.0867 (0.017)
	500 +	7127	300	0.01 (0.006)	0.0867 (0.017)
Employer-level	0 - 19	300	300	0.0133 (0.006)	0.153 (0.023)
	20 - 499	300	300	0.0033 (0.006)	0.0766 (0.017)
	500 +	301	300	0.0133 (0.006)	0.0800 (0.017)
Level	Firm Size Group	Observation	Unique SEIN	False Non-Match Rate Lower Bound	False Non-Match Rate Upper Bound
Establishment-level	0 - 19	300	300	0.44 (0.029)	0.667 (0.029)
	20 - 499	410	300	0.603 (0.029)	0.733 (0.029)
	500 +	2006	300	0.733 (0.029)	0.813 (0.029)
Employer-level	0 - 19	300	300	0.457 (0.029)	0.697 (0.029)
	20 - 499	478	300	0.67 (0.029)	0.82 (0.029)
	500 +	2813	300	0.767 (0.029)	0.86 (0.029)

Notes: Data are from the 2002 establishment-level and employer-level LEHD - BR crosswalk files, 2002 BR - SU and MU files, and 2002 quarter 1 - 4 ES - 202 files from the 2011 snapshot of the LEHD infrastructure files. An observation in the upper panel (false match rate) is a unique SEIN - FAS_EIN - State - BR establishment in the crosswalk files. An observation in the lower panel (false non-match rate) is a unique SEIN - SEINUNIT in the LEHD ES-202 file. The universe used in computing the false match rate and false non-match rate is 300 unique SEINs in each firm size category. The numerator used to calculate the lower bound of the false match (non-match) rate is the identified false match (non-matched) cases. The numerator used to calculate the upper bound of the false match (non-match) rate is the sum of the identified false match (non-match) cases and the unknown cases. Bootstrap standard errors are presented in parentheses.

Appendix Table 2.A. Match Rate at the FAS_EIN Level, By Year

Year	Establishment-Level Match			Employer-Level Match		
	Universe: (Valid/Active) FAS_EIN in					
		LEHD	BR		LEHD	BR
	Count	Percent	Percent	Count	Percent	Percent
	(1)	(2)	(3)	(4)	(5)	(6)
1990	2260000	88.7	40.3	2130000	83.4	40.0
1991	3100000	88.1	55.6	2940000	83.5	55.3
1992	3170000	87.8	56.3	2970000	82.1	55.9
1993	3340000	89.7	37.5	3140000	84.3	36.5
1994	3560000	86.4	61.0	3350000	81.3	60.7
1995	3940000	86.6	66.5	3720000	81.8	66.4
1996	4140000	84.9	69.1	3940000	80.8	69.0
1997	4220000	85.5	69.8	3990000	80.9	69.5
1998	4650000	83.3	76.5	4410000	78.9	76.4
1999	4790000	82.3	78.5	4550000	78.0	78.4
2000	5050000	83.7	82.2	4800000	79.6	82.2
2001	5120000	83.7	83.8	4900000	80.0	83.7
2002	5930000	93.3	66.5	5690000	89.4	65.6
2003	6110000	93.7	64.8	5890000	90.0	63.9
2004	6200000	93.9	62.1	5960000	90.3	61.2
2005	6330000	93.9	59.5	6100000	90.5	58.6
2006	6400000	93.9	57.0	6180000	90.7	56.2
2007	6460000	93.9	54.1	6210000	90.2	53.2
2008	6440000	93.6	51.7	6200000	90.1	50.8
2009	6300000	93.4	48.6	6070000	89.9	47.7
2010	6220000	93.0	46.3	6000000	89.6	45.4
2011	6170000	92.7	44.3	5950000	89.3	43.4
2012	5630000	93.3	38.7	5390000	89.3	37.7

Notes: Data from the 2011 LEHD Snapshot and 1990 - 2012 BR files are used. All the numbers are rounded to three significant digit for disclosure avoidance purposes. The universe used to compute the match rates in Columns (1) and (3) is valid FAS_EINs in LEHD by year. The universe used to compute the match rates in Columns (2) and (4) is EINs associated with active establishments in the BR files by year, after the duplicates are cleaned.

Chapter 3

The One Child Policy and Educational Attainment in China

3.1 Introduction

Controlling population growth and increasing child quality are generally considered critical for overall economic advancement at both the macroeconomic and the microeconomic levels. Among the policy instruments used in developing countries to promote economic development are programs designed to reduce the size of families (Rosenzweig and Zhang, 2009). One important motivation for these policies is the idea that reductions in family size can lead to increased allocation of resources for investments in human capital. This view is supported by the influential model developed by Becker and Lewis (1973), which suggests a trade-off between the quantity of children and average child quality. For policy makers in developing countries, understanding this trade-off is particularly relevant because many governments have attempted to reduce fertility with the explicit goal of increasing average population quality via human capital investment.

The empirical evidence on the magnitude of the quantity-quality trade-off, or even its existence, is mixed at best. On the one hand, some studies provide evidence showing that

family size has a negative effect on children's schooling (Blake, 1981; Hanushek, 1992; Lee, 2012; Rosenzweig and Wolpin, 1980; Rosenzweig and Zhang, 2009). On the other hand, more recent studies have found no effect, or even a positive effect, of family size on child outcomes (Angrist et al., 2005, 2010; Black et al., 2005; Guo and VanWey, 1999; Lee, 2008; Qian, 2009).

Theoretically, either outcome is possible within a consistent model of optimal family size and educational investment (Qian, 2009). For instance, the negative quantity-quality trade-off can arise from the assumption that the cost of average quality is an increasing function of the number of children (Becker and Lewis, 1973; Qian, 2009). Meanwhile, the complementarity between family size and children's average quality can arise from the assumption that there exist economies of scale in raising children. Though the assumption leading to a trade-off seems reasonable in many contexts, the economy of scale assumption may be equally reasonable in others. For example, transferring books, clothes, and knowledge is much easier among children of the same household than across households (Qian, 2009).

The challenge for empirical research has been to find sources of exogenous variation in family size that permit the identification of the trade-off between family size and average child quality as implied by the Becker-Lewis quantity-quality model (Angrist et al., 2010; Rosenzweig and Wolpin, 1980; Rosenzweig and Zhang, 2009; Qian, 2009). Direct interpretation and inference using the cross-sectional correlation between family size and the quality of children suffers from potential endogeneity bias. The main source of endogeneity is heterogeneity in parental preferences. For instance, one possibility could be that parents who value child quality more may also prefer to have fewer children. If such parental heterogeneity exists, the trade-off between family size and average child quality will be over-estimated. To address these concerns, existing studies have carefully designed identification strategies that exploit plausibly exogenous variation in the quantity of children caused by twinning or by the sex composition of the first two children (Angrist et al., 2005, 2010; Black et al., 2005; Conley, 2000; Conley and Glauber, 2006; Rosenzweig and Wolpin, 1980; Rosenzweig

and Zhang, 2009).

Rosenzweig and Zhang (2009), for example, study whether reduction in fertility increase human capital investment per child using the Chinese Child Twins Survey (CCTS) data collected in Yunnan Province in China. Because China's One Child Policy is strictly enforced in Yunnan's urban areas, twinning on the first birth results in one exogenous extra birth which allows the authors to identify the quantity-quality trade-off at birth parity one. Alternatively, in the rural areas where the survey was conducted, all families are permitted to have two children. The authors replicate the parity-N twins methodology at N = 2 to look at the impact of one extra birth on both the twins and the non-twins within the family. Overall, the authors find that an extra child at parity one or at parity two, net of one component of birth-endowment effects, significantly decreases the schooling progress, the expected college enrollment, and grades in school. Despite the significant trade-off between quantity of children and child quality in China, using the estimated effect of China's One Child Policy on family size from McElory and Yang (2000) and Qian (2009), Rosenzweig and Zhang (2009) claim that the contribution of the One Child Policy in China to the development of its human capital is modest.

Unlike most previous studies, which rely on the exogenous variation in family size provided by twinning, in this paper, the identification strategy is established by exploiting the exogenous variation provided by China's family planning policy and its subsequent relaxation to study child's quantity-quality trade-off in urban and rural China. Specifically, I partially address the issue of potential endogeneity in estimating the effect of family size on average child quality by instrumenting the family size with China's One Child Policy implementation and its subsequent relaxation.

Lee (2012) uses the China Health and Nutrition Survey data and shows that children in one-child households enjoyed significantly improved educational opportunities compared to children in multiple-child households. The analysis is done by gender group and for urban/rural households separately. The author attributes such an effect to China's One Child

Policy because he claims that without the policy, China would not have achieved such a high one-child rate so quickly (Lee, 2012). In addition, Lee (2012) also finds that the improvement was larger for girls than that for boys. To account for the potential endogeneity between family size and children's education, Lee (2012) instruments for family size (single-child versus multiple-child) using the height of a child. The author finds that the instrumental variable (IV) estimates and the ordinary least square estimates are similar. Though the author realizes the potential endogeneity between family size and children's education due to parents' tastes and preferences, the instrument chosen is poor and is very likely to have violated the exclusion restriction assumption of the Instrumental Variable approach. For instance, the height of a child may reflect his/her birth endowment that also has a positive effect on his/her schooling.

Qian (2009), on the other hand, exploits the plausible regional and time variation in the relaxation of China's One Child Policy and its effect on family size to estimate the trade off between family size and school enrollment of the first child in rural China. This paper matches the 1% sample of the 1990 Population Census with the 1989 China Health and Nutrition Survey at the county level and shows that an additional child significantly increased the probability of the school enrollment of first-born children by about 16 percentage points. In addition, Qian (2009) finds that such effect is larger for households where the children are of the same sex. Qian (2009)'s empirical strategy exploits three facts: first, an individual is only affected by the One Child Policy relaxation if she is born in a relaxed area; second, among first-born children born in relaxed areas, only girls are affected; and lastly, a girl is more likely to gain a sibling due to the policy relaxation if she is younger at the time of the policy relaxation enactment (Qian, 2009). [1] Qian (2009), however, studies rural households only. More importantly, the outcome variable used in her paper only measures whether one is currently enrolled in school, which can not measure and capture the quality of children properly.

[1] Details on the China's One Child Policy relaxation are provided in Section 3.2.

The empirical strategy utilized in this paper follows closely to Qian (2009). Instead of purely focusing on rural households, I analyze urban and rural Chinese households separately. Moreover, instead of focusing on the probability of whether one is currently enrolled in school, the outcome variable used in this paper is completed years of schooling which can measure the quality of children more accurately and properly. In this paper, only Han Chinese households are studied. Ethnic minorities were never included. The reasons are two-fold: first, ethnic minorities are always exempted from the China's family planning policy and its subsequent policy modification; and more importantly, during the early market reform period in China which was also the period when China's One Child Policy was announced, educational disparities between ethnic minorities and Han Chinese were reinforced (Hannum, 2002; Hannum et al., 2008). Specifically, using the census data, (Hannum, 2002) shows that between 1982 and 1990, although entrance rate to primary schools rose for all ethnic groups, the already-disadvantaged minorities suffered from a relative decrease in the transition rate to junior high school from primary school. Therefore, the rising disparity in transitions to secondary schools has strengthened the historical educational disadvantage of the ethnic minorities concentrated in less-developed regions (Hannum, 2002; Hannum et al., 2008). Such exacerbating disparity renders ethnic minorities a poor control group for this paper.

In this paper, I exploit the plausible exogenous changes in family size caused by China's initial One Child Policy and its subsequent policy relaxation to estimate the causal effect of family size on educational attainment in China. Specifically, to date, urban Han Chinese households have been continuously subject to the strict One Child Policy (OCP). I use the timing of the implementation of the policy to instrument for the family size of children born in urban Han Chinese households. This instrumental variable strategy exploits the fact that a Han Chinese who was born after the OCP implementation is affected while one born before should not be. In other words, I use the intent-to-treat effect of China's OCP to instrument family sizes.

In comparison, while rural Han Chinese households were initially subject to the One

Child Policy, a subsequent policy relaxation allowed rural families to have a second child if the first birth was a girl. The official relaxation occurred approximately five years after the initial policy implementation. Given the fairly short policy spacing, it is almost impossible to disentangle the effects on family size of the two. Therefore, for children born in rural areas, the policy variable used to instrument the family size will identify the composite effect of both the initial policy implementation and the subsequent relaxation. Two facts are exploited: first, an individual who was born after the implementation is affected by the policy itself and by the subsequent policy relaxation; second, family size is likely to be bigger in households whose first birth was a girl. The instrument for family size is therefore the interaction of the first-born child's sex and the One Child Policy variable. However, given the nature of the policy relaxation, parents may have chosen to keep girls in order to have a second child. Such household preferences could contaminate the instrumental variable estimates. This is the source of endogeneity that is not well tackled in this paper.

Overall, I find that the average family size has decreased substantially since the One Child Policy implementation. Additionally, first-born child being female in the rural area after the implementation of the One Child Policy has a positive effect on family size, which reflects the nature of the family planning policy relaxation and the "1-son-2-child" rule. For the urban-area sample, I find clear evidence indicating that there is indeed a negative trade-off between child's quantity and quality. Other things equal, an additional child can lead to a decrease of 1.2 years of schooling. A simple back-of-the-envelope calculation reveals that the implementation of the OCP has significantly increased the average completed years of schooling by approximately 0.68 years in urban China. A further comparison between urban female sub-sample and urban male sub-sample shows that, though the family policy implementation has a smaller effect on family size for female, the negative trade-off effect is much more prominent for female. It turns out that the implementation of the OCP has increased the average completed years of schooling for females more than that for males. However, unlike the urban-area sample, no significant effects of family size on educational

attainment for the rural sample are found.

There are caveats in this paper. For instance, the well-documented birth order effect and sibling sex composition effect are not analyzed here. Additionally, some potential endogeneity issues induced by the implementation of the OCP and its subsequent policy relaxation on rural households' preferences over girls are not well dealt with and this could contaminate the results in this paper. Thus, future research is needed to tackle these issues.

The paper is organized as follows. Section 3.2 review China's family planning policies. Section 3.3 reviews Becker and Lewis (1973) child quantity-quality trade-off model. Section 3.4 discusses the China Health and Nutrition Survey data and the sample used in this paper. Section 3.5 lays out the empirical identification strategy. Section 3.6 presents the descriptive statistics and econometric evidence. Finally, section 3.7 concludes.

3.2 Background

After two decades of rapid population growth under Mao's "people are wealth" propaganda, in the 1970s, the Chinese government enacted a serious of policies to slow population growth. Beginning in the early 1970s, the central government initiated the "Later [age], longer [the spacing of births], fewer [number of children]" family-planning campaign (McElory and Yang, 2000). Implementation of the campaign relied mostly on propaganda, "persuasion,", and social pressure (McElory and Yang, 2000). Incentives in various forms for parents who spaced the births of their children at least four years apart was offered (Croll et al., 1985).

However, due to the rapid population growth in the 1950s and 1960s, the age structure of China's population by the end of 1970s was such that 50% and 65% of the population were under the ages of 20 and 30, respectively (Croll et al., 1985). Absent aggressive abatement strategies, population growth rates were expected to be extraordinarily high during the 1980s. As a consequence, the government abandoned these indirect controls and moved to a more restrictive policy, which is the so-called the One Child Policy (OCP) (Croll et al., 1985;

Edlund et al., 2007; McElory and Yang, 2000; Rosenzweig and Zhang, 2009; Lee, 2012). The OCP was officially announced in 1979. Although it is recorded that the actual implementation began in certain region as early as 1978, the One Child Policy was firmly enforced by 1980 (Croll et al., 1985). The policy applies to all individuals of Han Chinese ethnicity, a group that comprises more than 90% of China's total population. Ethnic minorities were always exempted from the family planning policies. The specifics of the OCP and its enforcement vary from one place to another in China (Rosenzweig and Zhang, 2009). Usually, in urban areas, the OCP was strictly enforced and second births were permitted only in a handful of extenuating circumstances (McElory and Yang, 2000; Rosenzweig and Zhang, 2009). In rural areas, however, the OCP was enforced in a laxer manner. For instance, (Rosenzweig and Zhang, 2009) find that households from some rural communities in Yunnan Province are encouraged to have one child but are always exempted from the strict OCP.

Unexpectedly, China's family planning policy led to massive female infanticide, gender-specific abortions, and civil protests upon its implementation. These incidents were particularly pervasive in the rural areas where the long-existing son preference is much stronger. As a results, in rural areas, the One Child Policies were liberalized (McElory and Yang, 2000; Greenlaugh, 1986; Qian, 2009). As shown in civil records, local governments began issuing permits for a second child as early as 1982. In response to severe political pressure and to curb female infanticide, in 1984 the central government officially issued "Document 7," which introduced the policy relaxation to the OCP. Although this policy reform still strictly prohibited third births among Han Chinese, it significantly expanded the conditions under which couples were allowed to have second births. The main relaxation following "Document 7" is called the "1-son-2-child" rule (Greenlaugh, 1986). This rule allowed rural Han Chinese couples to have a second child if their first born child was a girl (Greenlaugh, 1986). Qian (2009) cites statistics showing that while in 1982 the second child permits were allotted to only 5% of rural households, this share had increased to 50% by 1986.

Given the differential enforcement of the One Child Policy and the subsequent rural-

area specific One Child Policy relaxation, this paper will separately analyze urban and rural households. In addition, because of the close spacing between the initial OCP announcement and its subsequent relaxation, the regional and temporal variation in the announcement of the policy relaxation, and the variation in the enforcement across rural communities, I use the combined intent-to-treat effect of China's OCP and its subsequent relaxation to instrument family size for rural-area households.

3.3 Conceptual Framework

In this section, I present a simple model of households that allows for heterogeneity in child quality, which closely follows the work by (Becker and Lewis, 1973; Becker and Tomes, 1976; Qian, 2009). I assume that each household has a utility function of the following general form

$$U = U(n, \{e_i\}_{i=1}^{n}, y) \tag{3.1}$$

where n is the number of children, $\{e_i\}_{i=1}^{n}$ represents the quality of each child in terms of educational attainment, and y is the aggregate amount of all other commodities. The cost of achieving education level e_i depends on the level of education itself and the number of children in the household. The function can be represented as

$$c(e_i) = h(e_i, n) \tag{3.2}$$

where $c(e_i)$ is a convex function in e_i. If there is a fixed cost associated with educating each child, which is denoted as z, then the total cost of education in a household with n children and chosen levels of education $\{e_i\}_{i=1}^{n}$ can be written as

$$\sum_{i=1}^{n}(z + h(e_i, n)).$$

Denote the price level of the other commodities as p. Then, the budget constraint faced by a household is

$$\sum_{i=1}^{n}(z + h(e_i, n)) + py \leqslant I \tag{3.3}$$

where I is the total household income.

The constrained maximization problem faced by this household can be written as

$$\varsigma = maxU(n, \{e_i\}_{i=1}^{n}, y) - \lambda[\sum_{i=1}^{n}(z + h(e_i, n)) + py - I]. \tag{3.4}$$

The first order conditions are

$$\frac{\partial \varsigma}{\partial n} = \frac{\partial U}{\partial n} - \lambda[z + h(e_i, n) + nh_n(e_i, n)] \equiv 0 \tag{3.5}$$

$$\frac{\partial \varsigma}{\partial e_i} = \frac{\partial U}{\partial e_i} - \lambda[nh_{e_i}(e_i, n)] \equiv 0$$

$$\frac{\partial \varsigma}{\partial y} = \frac{\partial U}{\partial y} - \lambda p \equiv 0.$$

Denote the optimal education, number of children, and other commodity consumption as e^*, n^*, and y^*. At the optimum, for all i,

$$\frac{\partial U}{\partial n^*} = \lambda[z + h(e^*, n^*) + n^* h_{n^*}(e^*, n^*)] \tag{3.6}$$

$$\frac{\partial U}{\partial e^*} = \lambda[n^* h_{e^*}(e^*, n^*)]$$

$$\frac{\partial U}{\partial y^*} = \lambda p.$$

The negative trade-off between the quantity and quality of children stems from the fact that the shadow price of children with respect to quantity is positively correlated with the quality of children. Similarly, the shadow price of children with respect to quality is positively correlated with the quantity of children. For instance, at the optimum, any positive perturbation in n^* will increase the shadow price of children with respect to quality,

which will lead to a decrease in e. And vise versa, any positive perturbation in e^* will increase in the shadow price of children with respect to quantity, which will lead to a decrease in n.

3.4 Data & Sample

3.4.1 Data

The data used in this paper are from the 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 *China Health and Nutrition Survey* (CHNS). The survey uses a multistage, random cluster process to draw a sample of about 4,400 households with a total of 19,000 individuals in nine provinces that vary substantially in geography, economic development, and other critical variables. [2] Specifically, counties in the nine provinces were first stratified by income, and a weighted sampling scheme was used to randomly select four counties in each province. In addition, the provincial capital and a lower income city were selected when feasible. Villages and townships within the counties and urban and suburban neighborhoods within the cities were selected randomly. In 1989 to 1993 there were 190 primary sampling units, and a new province and its sampling units were added in 1997.

According to the CHNS research team, the follow-up levels are high and the follow-ups of households that move within the primary sampling units and some larger urban entities are attempted. However, families that migrate from one community to a new one (a non-primary-sample-unit community) are not followed. The first round of the CHNS, including household, community, and health/family planning facility data, was collected in 1989. Since the 1993 survey, all new households formed from sample households were added. Since 1997, new households in original communities were also added to replace households no longer participating in the study. Also since 1997, new communities in original provinces have been added to replace sites no longer participating. A new province was also added in 1997 when

[2]These nine provinces are: Guangxi, Guizhou, Heilongjiang, Henan, Hubei, Hunan, Jiangsu, Liaoning, and Shandong.

one province was unable to participate. The dropped province returned to the study in 2000. [3]

The 1989 CHNS surveyed 15,917 individuals comprised 3795 households. The 1991 CHNS surveyed only individuals belonging to the original sample households, which resulted in a total of 14,778 individuals comprised 3616 households. For the 1993 CHNS, all new households formed from sample households who resided in sample areas were added to this sample, resulting in a total of 13,893 individuals comprised of 3441 households. For the CHNS 1997, all newly formed households who resided in sample areas (and additional households to replace those no longer participating) were added to the sample. New communities were also added to replace communities no longer participating, and Heilongjiang province replaced Liaoning province. A total of 14,426 individuals participated in 1997 comprised of 3875 households. In the 2000 CHNS, newly formed households, replacement households, and replacement communities were again added, and Liaoning province returned to the study. A total of 15,648 individuals participated in 2000 comprised of 4403 households. [4]

The CHNS data include variables for educational attainment measured in completed years of schooling, relationships among household members, and many other individual-, household-, and community-level characteristics. In particular, the Ever-Married Women Survey component includes complex marriage and fertility histories of women who are 52 years old or younger. This survey component allows children to be linked to their mothers within the household. Thus, family size can be calculated. Using birth date, I can reconstruct birth order and family composition in the sample. Births outside of marriage were rare in China during this period. Therefore, the family size calculation is considered to be rather reliable.

[3] Details are obtained from `http://www.cpc.unc.edu/projects/china/about/proj_desc/survey`.

[4] The sample description is only available for the 1989, 1991, 1993, 1997, and 2000 CHNS. It can be found at `http://www.cpc.unc.edu/projects/china/about/design/sample`.

3.4.2 Sample

As discussed before, the samples used in this paper only include Han Chinese individuals. Given the high follow-up level, for an individual who is interviewed in more than one wave of the CHNS survey, data from the most recent survey year is utilized for this individual. If the educational attainment measured in completed years of schooling is missing in the most recent survey year, the highest non-missing educational attainment variable from previous survey years is used.

Unlike most literature on U.S. education which usually uses population aged 25 and over as target sample group, the urban and rural samples in this paper are restricted to individuals who are at least 18 years older and currently not in school. The main reason is two-fold: first, China has a nine-year compulsory education system, which includes six years of primary education starting at age six or seven, and three years of junior secondary education covering ages 12-15(or 13-16 if one starts primary school at age seven); [5] second, the average education level in China is much lower and most people in China have finished school by 18 or even younger. Table 3.1 provides the educational attainment for total population aged 15 and over in China from 1950 to 2000. The data is borrowed from the Barro-Lee Educational Attainment Dataset. [6] From Table 3.1 we can see that from 1950 to 2010, the average years of schooling have more than tripled for the population aged 15 and over. Even so, in 2010, the average years of schooling is only 8.11 years. That is to say, in 2010 among population aged 15 and over, an average educated person would have completed school when she was 14 or 15 years old (depending on whether she started at age 6 or 7). Hence, by focusing on individuals aged 18 and over in the CHNS data, I should be able to capture most individuals' completed education levels.

The final urban-area samples include 951 individuals with 436 females and 515 males. The final rural-area samples include 2158 individuals with 923 females and 1235 males.

[5] Details on the education system can be found on China's Ministry of Education website.

[6] Details about the Barro-Lee Educational Attainment Dataset can be found at `http://www.barrolee.com/` and Barro and Lee (2013).

3.5 Empirical Strategy

I conduct empirical analyses and estimation for urban and rural households separately and contrast the findings. The rationale for separate urban and rural analyses are: first, urban and rural households endure differential treatment from China's family planning policies. Overall, urban households have been continuously subjected to the strict One Child Policy whereas rural households were initially subject to the strict One Child Policy and later became eligible for the policy relaxation. Second, urban and rural households may have heterogeneous preferences over children quantity, quality, and other demographic characteristics. The One Child Policy relaxation for rural households is believed to be a direct consequence of female infanticide, pervasive sex-selective abortions, and civil protests in rural China (Croll et al., 1985). Overall, rural households may have stronger son preferences due to social norms and economic exigencies. Such intrinsic preference heterogeneity makes it necessary to analyze rural and urban households separately. Third, China has always been characterized by its urban-rural divide in a wide range of economic aspects such as income, education, and health. Since the economic reform was initiated, that urban-rural divide has widened substantially (Knight and Song, 1999). Inequality in education provision between urban and rural areas significantly increased in the 1980s due to the decentralization of expenditure responsibilities and the reduction in subsidies from rich regions to poor regions. Hannum and Park (2007) find that many rural schools were closed and rural school enrollment rates decreased substantially. It is also likely that the unobservable quality of schools declined due to lack of funding in rural areas relative to urban areas. To allow for different trends related to these factors, I conduct analyses for rural and urban households separately.

The main structural equation can be written as

$$schooling_{itc,y} = \alpha + sibsize_{itc,y}\beta + X'_{itc,y}\delta + ttrend_y\phi + \gamma_c + \epsilon_{itc,y} \tag{3.7}$$

where the dependent variable $schooling_{itc,y}$ measures completed years of formal schooling for individual i, in province c, born in cohort t (t can take the following two values: pre-OCP or post-OCP), observed in CHNS wave y; [7] $sibsize_{itc,y}$ measures the family size; $X_{itc,y}$ is a vector of individual characteristics which is comprised of individual i's sex, mother's age at birth, mother's education, and log per capita household income. Lastly, γ_c controls for fixed province effects. $ttrend_y$ controls for a linear time trend where $ttrend_y = y - 1988$. [8] $\epsilon_{itc,y}$ is the error term.

This structural equation faces the potential endogeneity problem that family size and human capital investment in these children are jointly determined by parents. For instance, one possibility is that parents who place higher value on education also prefer fewer children. If this were the case, the simple least squares estimates will over-estimate the (absolute value of the) negative effect of an additional child on schooling. To address this concern, I exploit the plausibly exogenous variation in family size provided by the One Child Policy and its subsequent policy relaxation.

Because urban households were strictly subject to the One Child Policy without relaxation, regardless the sex of children, family size should be negatively correlated with year of birth. In other words, year of birth determines an individual's exposure to the implementation of the One Child Policy, which is exogenous. The exclusion restriction for the instrument is that it must be correlated with family size and affect the outcome variable, years of schooling, only through *sibsize*. The identification strategy for urban households

[7] Recall, as described in Section 1.3.2, for an individual who is interviewed in more than one wave of the CHNS survey, data from the most recent survey year is utilized for this individual.

[8] For the urban sample, an alternative specification with a fixed survey-year effect rather than a linear time trend is also explored. The results from both specifications are very similar. However, given the sample size, the alternative specification with the fixed survey-year effect is not of sufficient rank to perform model tests when I estimate Equation 3.7 for female and male separately. Additionally, the fixed survey-year effect and province effect is further replaced by fixed province trends. Similarly, given the sample size, it is not of sufficient rank to perform the model test. Therefore, for this paper, the preferred specification is the one controlling for a linear time trend and fixed province effect. For the rural sample, similar conclusions can be reached regarding the specifications with the fixed survey-wave effect and with the linear time trend. The rural sample size is large enough to allow specification with fixed province trends. The results, which are not reported in the paper, are very comparable to those obtained using the preferred specification.

can be written as

$$sibsize_{itc,y} = \tau + OCP_{itc,y}\theta + X'_{itc,y}\eta + ttrend_y\phi + \gamma_c + \xi_{itc,y} \tag{3.8}$$

Since the official announcement date of the OCP was 1979, I set $OCP_{itc,y}$ equal to one for individuals born in or after 1979 and zero otherwise. All other covariates are defined as in Equation 3.7.

In comparison, rural households were subject to the initial One Child Policy just like urban households. However, approximately three to four years into the policy, rural households became eligible for the One Child Policy relaxation, which is often referred to as the "1-son-2-child" rule. The relaxation allowed rural parents to have a second child if the first-born child was a girl. Therefore, year of birth and the sex of a household's first-born child jointly determine an individual's exposure to the policy relaxation. First, family size should be negatively correlated with exposure to the initial OCP implementation. Among those individuals who were born after the OCP implementation, given the "1-son-2-child" rule, family size should be positively correlated with the first-born being a girl. The interaction term of the first-born's sex and one's year of birth therefore, captures the differential effect of the relaxation on family size. To summarize, the instruments for family size for rural households include the following two terms: an individual's year of birth, which determines her exposure to the initial OCP implementation; and the interaction of the first-born's sex with one's year of birth, which captures the additional effect due to the subsequent OCP relaxation. The identification strategy for rural households can be written as

$$sibsize_{itc,t} = \tau + OCPRelax_{itc,y}\theta + OCPRelax_{itc,y} \times FB\_sex_{itc,y}\psi + X'_{itc,y}\eta + ttrend_y\phi + \gamma_c + \xi_{itc,y} \tag{3.9}$$

As noted above, $OCPRelax_{itc,y}$ equals one for individuals born in or after 1979 and zero otherwise. It captures the composite effect due to the initial OCP implementation and its subsequent modification. In Equation 3.9, FB_sex is an indicator for first-born being

a girl. Therefore, the endogenous family size is instrumented using $OCPRelax_{itc,y}$ and $OCPRelax_{itc,y} \times FB\_sex_{itc,y}$. All other covariates are defined as in Equation 3.7. In order to accurately estimate the main effect and the interaction effect, for rural households, $FB\_sex_{itc,y}$ is added to the vector of covariates $X_{itc,y}$ in both the first-stage and the second-stage equations.

3.6 Empirical Results

3.6.1 Descriptive Statistics

Figure 3.1 and Figure 3.2 plot the share of male and female in the urban-area sample and the rural-area sample by the One Child Policy variable, respectively. The One Child Policy variable for the urban-area sample is referred to as "OCP" mainly because urban households are only subjected to the implementation of the original One Child Policy. Individuals born in or before 1978 are considered to the pre-OCP cohort and individuals born in or after 1979 are considered to be the post-OCP cohort. The One Child Policy variable for the rural-area sample is referred to as "OCPRelax" to acknowledge the combined effect from both the initial One Child Policy implementation and its subsequent policy relaxation. As discussed before, given the close spacing between the initial implementation and the subsequent relaxation, as well as the differential enforcement across localities, disentangling the two effects is very difficult. Individuals born in or before 1978 in rural areas are considered to be the pre-OCPRelax cohort and individuals born in or after 1979 are considered to be the post-OCPRelax cohort.

Specifically, Figure 3.1 shows that the post-OCP cohorts in the urban areas clearly have more unbalanced sex ratios (male/female) as compared to the pre-OCP cohorts. A similar pattern exists in the rural sample as well, which can be seen in Figure 3.2. Using China's 1990 Census data and a difference-in-difference estimator, Zhang (2011) find that the enforcement of the OCP has led to 4.4 extra boys per 100 girls in the 1980s, which can account for about

94% of the total increase in sex ratios during this period. Figure 3.1 and Figure 3.2 confirm the results in Zhang (2011).

Table 3.2 presents the descriptive statistics by birth cohort (pre-OCP cohort vs post-OCP cohort) and sex using the urban sample. A brief comparison reveals that, the post-OCP individuals, on average, live in households with higher per capita income (Log Per Capita Household Income) and are younger (Age) compared to the pre-OCP individuals. The mothers of the post-OCP individuals are on averaged better educated (Mother's Education) and share similar age when giving birth (Mother's Age at Birth).

Table 3.2 also shows that the average years of schooling have increased comparing the pre-OCP and the post-OCP cohorts. For the pre-OCP cohort (OCP = 0), the average completed years of schooling is 10.4 years. In comparison, the post-OCP cohort, on average, has 12.4 years of schooling. The magnitude of the increase is larger for females than for males. A direct consequence is a reversal in the educational gap between men and women. In particular, the average years of schooling for men born in urban areas before the implementation of the OCP is around 10.6 years, whereas the average number of years of schooling for women born in the same cohort is slightly less 10.3 years. This difference disappears when we look at men and women born after the OCP implementation. In fact, among the post-OCP cohorts, women have slightly higher education than men (12.7 years for women and 12.2 for men). The trend in average completed years of schooling by birth cohort and sex for the urban sample is further visualized in Figure 3.3.

Table 3.2 also shows that the average family size (number of children which is measured by sibsize) has decreased substantially since the OCP implementation. Specifically, for an average individual in pre-OCP cohort, the average family size was 2.69. This number decreased by almost one birth to 1.75 for an average post-OCP-cohort individual. This statement holds true for both men and women, although the decrease in average family size for women is smaller than that for men. On average, women always live in households with more children. The trend in family size before and after the implementation of the One

Child Policy is illustrated in Figure 3.5. Taken together, the descriptive statistics computed using the sample of urban households suggest a negative correlation between family size and education, which suggests that there is a trade-off between the quantity and quality of children.

Table 3.3 computes the same set of descriptive statistics for the rural sample. In addition, Figure 3.4 and Figure 3.6 depict similar trends in mean family size and completed years of schooling for the rural sample. The overall trends described above for the urban sample also hold for the rural sample. From Table 3.3 one can see that a trade-off between the quantity and quality of children also seems to exist. The negative trade-off seems to be more prominent for women. In addition, a close comparison between Table 3.2 and Table 3.3 shows a smaller effect of China's family planning policy on family size (sibsize) in rural households compared to urban households. It seems to confirm the existence of an effect of both the OCP implementation and its subsequent policy relaxation. Given the nature of the "Document 7" and the "1-son-2-child" rule (Qian, 2009), among individuals who were born in or after 1979, family size should be positively correlated with the first-born child being a girl. Table 3.4 presents the descriptive statistics for the post-OCPRelax cohort by first-born child sex. From Table 3.4, we can clearly see that, for the post-OCPRelax cohort, average family size of those households with female first-borns is substantially larger (2.83 vs 2.30). Table 3.4 further confirms the positive effect of the OCP relaxation on family size in rural China.

In summary, the descriptive statistics presented thus far suggest that China's family planning policies do have a significant effect on family size and that there seems to be a negative trade-off between family size and child's educational attainment. In Section 3.6.2, econometric evidence is presented to examine whether indeed there is a causal relationship between family size and educational attainment.

3.6.2 Econometric Evidence

Table 3.5 shows the simple least squares estimates and instrumental variables (2SLS) estimates of the quantity-quality trade-off effect for the urban sample. This is estimated using Equation 3.7 and Equation 3.8. The vector of individual characteristics includes individual i's sex, mother's age at birth, mother's education, and log per capita household income. Results from two specifications are reported: one with linear time trends and one with survey-wave fixed effects. The preferred specification is the one controlling for the linear time trends (Columns (4) - (6) in Table 3.5). Results from these three columns will be discussed.

The upper panel in Table 3.5 presents the simple OLS estimates using the urban sample, the urban female sub-sample, and the urban male sub-sample, separately. The OLS estimates indicate that, other things equal, an additional sibling is associated with 0.47 years less schooling. This finding is consistent with the existence of trade-offs between the quantity and quality of children. When I repeat the same exercise using the urban female sub-sample and the urban male sub-sample separately the negative quantity-quality trade off appears larger for men than for women.

To address the potential endogeneity issue of family size as discussed before, the 2SLS estimate from the first-stage and second-stage are presented in the middle and lower panels of Table 3.5. The first-stage regression result indicates that the implementation of the One Child Policy had a significant effect on family size. Specifically, other things equal, the implementation of the One Child Policy directly led to a decrease in family size of 0.57 children for the urban sample. In addition, the main-stage estimate shows that an additional child can lead to a decrease of 1.2 years of schooling. A simple back-of-the-envelope calculation reveals that the implementation of the OCP has significantly increased the average completed years of schooling for urban sample by approximately 0.68 years (= (-0.57) $\times$ (-1.20)). I further repeat the 2SLS estimations for urban female and urban male sub-samples separately. The results show that the implementation of the OCP has a larger

effect on family size for males than for females (-0.75 for males and -0.27 for females). However, other things equal, an additional child can decrease the average years of schooling for females by 2.8 years, which is three times as large as the effect of one additional child on male educational attainment. A similar back-of-the-envelope calculation shows that the implementation of the OCP has increased the average completed years of schooling for female by about 0.76 years, which is larger than that for male (0.68 years). This may help explain the disappearance of educational gap between males and females before and after the OCP implementation.

Table 3.6 presents the simple OLS estimates and 2SLS estimates of the quantity-quality trade-off effect for rural households. This is done using Equation 3.7 and Equation 3.9. The vector of individual characteristics includes individual i's sex, mother's age at birth, mother's education, log per capita household income, and first-born child's sex. The last covariate is added such that the main effect and the interaction effect of the OCP implementation and its subsequent relaxation can be correctly estimated using the 2SLS estimates. This variable is also added to the OLS specification to ensure comparable results obtained using OLS and 2SLS estimation strategies. Similarly, results from two specifications are reported: one with linear time trends and one with survey-wave fixed effects. As discussed in Footnote 8, the preferred specification is the one controlling for the linear time trends (Columns (4) - (6) in Table 3.6). Results from these three columns will be discussed.

The upper panel in Table 3.6 presents the simple OLS estimates using the rural sample, the rural female sub-sample, and the rural male sub-sample, separately. Similar to the urban households, rural households also exhibited a negative trade-offs between the quantity and quality of children when using the full rural sample and the rural female sub-sample. When comparing results obtained using the female and male sub-sample, it appears that the negative trade-off effect between family size and educational attainment is larger and significant for female relative to male. This seems to suggest that the particularly strong, long-lasting, and persistent son preference in the rural area has rendered women in a disadvantaged po-

sition.

Due to concerns about potential endogeneity issues, I also present the 2SLS estimates using Equation 3.7 and Equation 3.9. To restate, the identification strategy exploits two facts here: first, individuals are exposed to the treatment of the initial OCP implementation and its subsequent relaxation if and only if one was born in or after 1979; secondly, among individuals who were born after the initial OCP implementation, one is only affected by the policy relaxation if the first-born child in his household is female. Therefore, for rural sample, family size is instrumented using the policy variable (OCPRelax) and the interaction term between OCPRelax and the first-born child's sex (FB_sex). The first-stage and second-stage results are presented in Table 3.6.

The lower panel in Table 3.6 shows the first-stage result: the combined effect of the OCP implementation and its subsequent policy relaxation on rural family size. The first-stage result first exhibits a clear binding effect due to the initial implementation and the subsequent policy relaxation. For instance, the OCP implementation and its subsequent relaxation have significantly decreased the average family size by 0.44 births (Column (4) in Table 3.6. It is evaluated at the mean first-born sex (which is 0.48 in the full rural sample), i.e., $-0.60 + 0.33 \times 0.48$.). Here, we can see that the effect of the family policy on family size in rural China is noticeably smaller in comparison to urban China. Overall, this similar effect and trend holds true for both the rural female sub-sample and the rural male sub-sample. As hypothesized, the subsequent policy relaxation, the "1-son-2-child," rule, has a significant positive effect on family size for households with first-born child being female. Other things equal, individuals born in or after 1979 to households with first-born child being female, on average, live in households with 0.33 more children compared to individuals born in the same cohorts but to households with first-born child being male. Interestingly, when I examine the female sub-sample and the male sub-sample separately, the interaction effect is only significant for female group.

However, unlike the urban-area sample, the second-stage estimate of the effect of family

size on educational attainment for the rural sample are all insignificant. On possible cause for the results is that there is an additional potential source of endogeneity is not well accounted for in this paper. For instance, when the One Child Policy relaxation is announced, rural parents may choose to keep girls so that they can be eligible for the "1-son-2-child" rule. Therefore, the 2SLS estimates presented here may be biased by the change in parental preferences and they should be interpreted with caution.

3.7 Conclusion

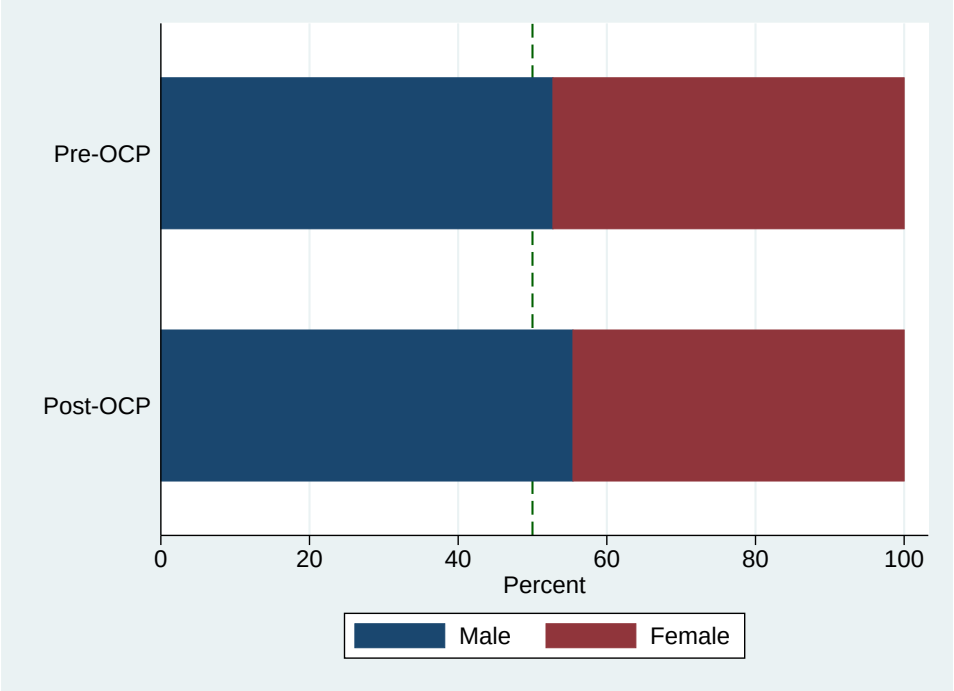
Many believe that family size and child's average quality are inversely correlated. This paper exploits plausibly exogenous changes in family size caused by the initial implementation and subsequent relaxations in China's One Child Policy to estimate the causal effect of family size on educational attainment. The data used in this paper are from the 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 *China Health and Nutrition Survey* (CHNS).

Overall, I find that the average family size has decreased substantially since the One Child Policy implementation, which implies the binding effect of China's family planning policy. Additionally, first-born child being female in the rural area after the implementation of the One Child Policy has a positive effect on family size, which reflects the nature of the family planning policy relaxation and the "1-son-2-child" rule. For the urban-area sample, I find clear evidence indicating that there is indeed a trade-off between child's quantity and quality. Other things equal, an additional child can lead to a decrease of 1.2 years of schooling. A simple back-of-the-envelope calculation reveals that the implementation of the OCP has significantly increased the average completed years of schooling of urban sample by approximately 0.68 years. A further comparison between female sub-sample and male sub-sample shows that, though the family policy implementation has a smaller effect on family size for female, the negative trade-off effect is much more prominent for female. Thus, it turns out that the implementation of the OCP has increased female average completed years

of schooling more than that for male by 0.076 years.

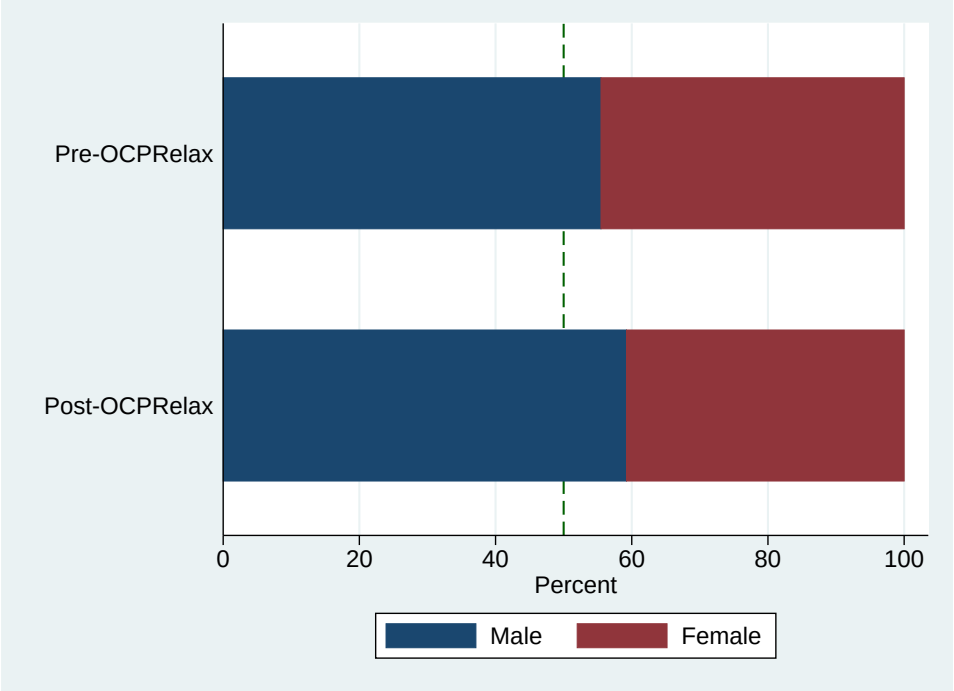
Unlike the urban-area sample, no significant effects of family size on educational attainment for the rural sample are found. One possible reason for the results is other potential sources of endogeneity that is not well accounted for. For instance, when the One Child Policy relaxation is announced, rural parents may choose to keep girls so that they can be eligible for the "1-son-2-child" rule. Therefore, the 2SLS estimates presented here may be biased due to the parental preference change and should be interpreted with caution. Several other caveats also exist, for instance, the well-documented birth order effect and sibling sex composition effect are overlooked in this paper. Cautious interpretation of the results is recommended.

Figure 3.1. Percentage of Male and Female By One Child Policy Variable (OCP) in Urban Area



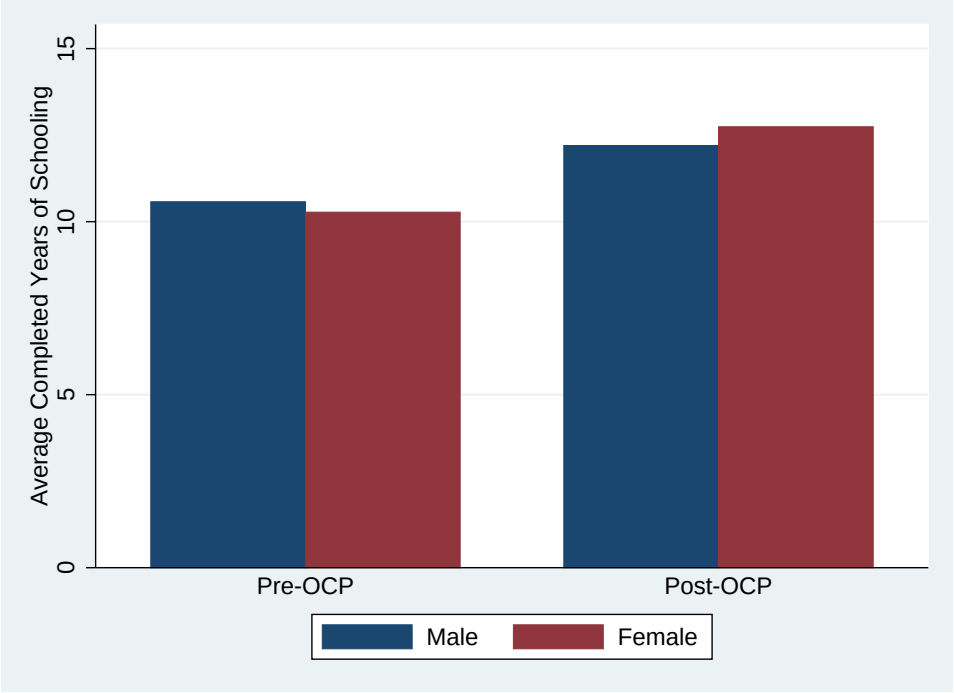
Notes: Data are based on author's own calculation using the China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data. The urban-area sample includes individuals who are at least 18 years old and currently not in school. Pre-OCP cohort includes individuals born before 1979; Post-OCP cohort includes individuals born in or after 1979. Blue bars represent male. Red bars represent female. The green dash line is the reference line at 50 percentage point.

Figure 3.2. Percentage of Male and Female By One Child Policy Variable (OCPRelax) in Rural Area



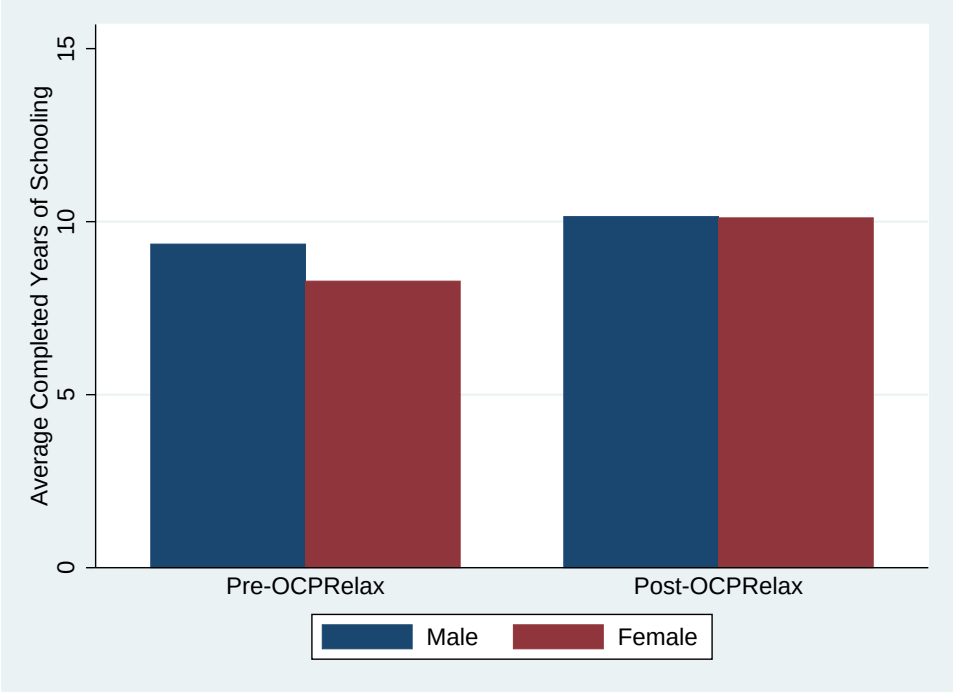
Notes: Data are based on author's own calculation using the China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data. The rural-area sample includes individuals who are at least 18 years old and currently not in school. Pre-OCPRelax cohort includes individuals born before 1979; Post-OCPRelax cohort includes individuals born in or after 1979. Blue bars represent male. Red bars represent female. The green dash line is the reference line at 50 percentage point.

Figure 3.3. Average Completed Years of Schooling By Sex and One Child Policy Variable (OCP) in Urban Area



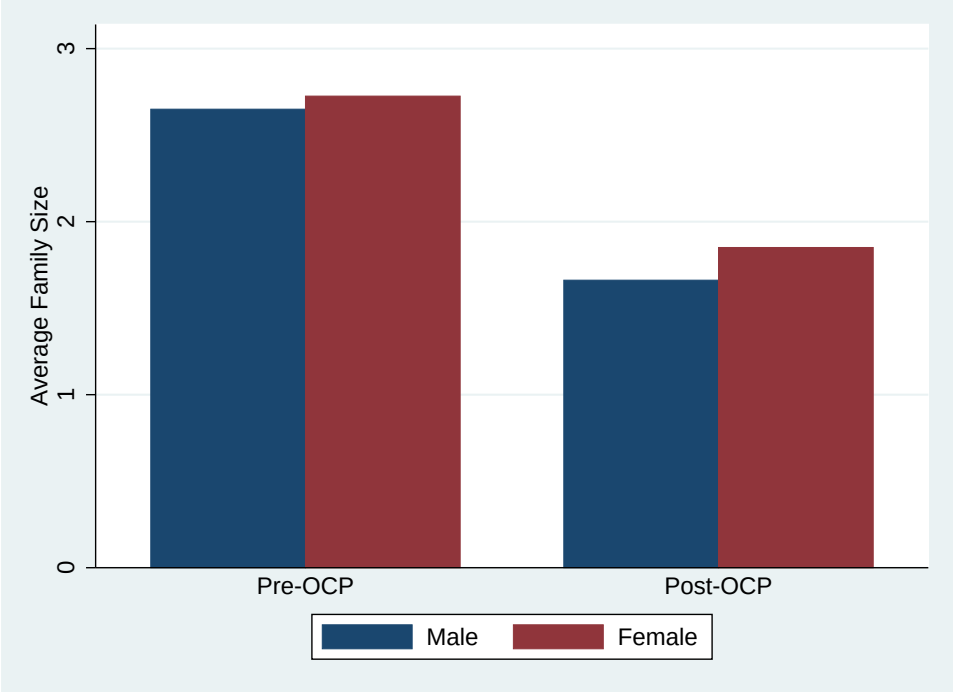
Notes: Data are based on author's own calculation using the China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data. The urban-area sample includes individuals who are at least 18 years old and currently not in school. Pre-OCP cohort includes individuals born before 1979; Post-OCP cohort includes individuals born in or after 1979. Blue bars represent male. Red bars represent female.

Figure 3.4. Average Completed Years of Schooling By Sex and One Child Policy Variable (OCPRelax) in Rural Area



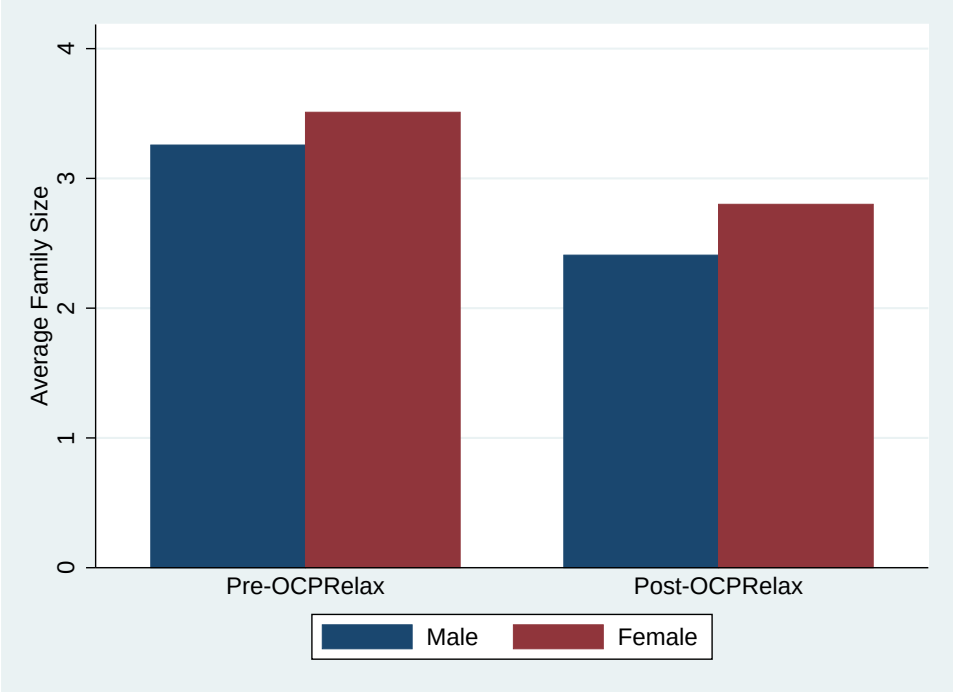
Notes: Data are based on author's own calculation using the China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data. The rural-area sample includes individuals who are at least 18 years old and currently not in school. Pre-OCPRelax cohort includes individuals born before 1979; Post-OCPRelax cohort includes individuals born in or after 1979. Blue bars represent male. Red bars represent female.

Figure 3.5. Average Family Size By Sex and One Child Policy Variable (OCP) in Urban Area



Notes: Data are based on author's own calculation using the China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data. The urban-area sample includes individuals who are at least 18 years old and currently not in school. Pre-OCP cohort includes individuals born before 1979; Post-OCP cohort includes individuals born in or after 1979. Blue bars represent male. Red bars represent female.

Figure 3.6. Average Family Size By Sex and One Child Policy Variable (OCPRelax) in Rural Area



Notes: Data are based on author's own calculation using the China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data. The rural-area sample includes individuals who are at least 18 years old and currently not in school. Pre-OCPRelax cohort includes individuals born before 1979; Post-OCPRelax cohort includes individuals born in or after 1979. Blue bars represent male. Red bars represent female.

Table 3.1. Educational Attainment for Total Population Aged 15 and Over in China, 1950-2010

Year	Avg. Years of Total Schooling	Highest Level Attained					
		Primary		Secondary		Tertiary	
		Total	Completed	Total	Completed	Total	Completed
		% of population aged 15 and over					
1950	1.57	21.79	6.79	8.01	1.47	0.34	0.17
1955	1.86	25.31	8.17	9.44	1.82	0.49	0.25
1960	2.34	28.55	12.04	12.30	2.48	0.68	0.36
1965	2.78	32.74	12.58	15.51	3.45	0.84	0.45
1970	3.43	36.93	16.66	20.28	4.58	0.84	0.43
1975	3.97	38.46	18.69	25.24	5.87	0.88	0.44
1980	4.75	38.35	20.07	33.64	9.34	0.91	0.45
1985	5.25	36.95	20.51	37.96	14.55	1.50	0.78
1990	5.62	34.55	19.89	41.30	21.78	1.93	1.02
1995	6.41	32.95	19.41	47.85	29.57	3.26	1.80
2000	7.11	30.38	18.33	54.06	36.24	4.59	2.59
2005	7.60	27.65	17.07	57.83	41.64	6.09	3.42
2010	8.11	24.45	15.28	60.95	46.49	8.05	4.58

Data Source: Education Attainment for Population Aged 15 and Over (`http://www.barrolee.com/data/full1.htm`).

Table 3.2. Descriptive Statistics: By One Child Policy Variable (OCP) and Sex, Urban Area

Variable	Summary Statistics					
	OCP = 0			OCP = 1		
	All	Female	Male	All	Female	Male
School	10.4	10.3	10.6	12.4	12.7	12.2
	(2.72)	(2.72)	(2.73)	(2.76)	(2.90)	(2.61)
Sibsize	2.69	2.73	2.65	1.75	1.85	1.67
	(1.26)	(1.24)	(1.27)	(0.89)	(1.01)	(0.78)
Mother's Education	4.91	4.72	5.08	8.52	8.33	8.67
	(3.98)	(4.15)	(3.83)	(3.60)	(3.55)	(3.65)
Mother's Age at Birth	26.74	27.03	26.5	25.7	25.6	25.8
	(5.10)	(5.08)	(5.12)	(3.65)	(3.64)	(3.66)
Log Per Capita Household Income	8.67	8.51	8.82	9.25	9.19	9.30
	(0.98)	(0.87)	(1.04)	(1.05)	(1.04)	(1.07)
Age	28.1	25.3	30.6	24.3	23.8	24.7
	(6.96)	(5.55)	(7.14)	(3.47)	(3.33)	(3.54)
N	455	215	240	496	221	275

Notes: The China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data are used. The urban-area sample includes individuals who are at least 18 years old and currently not in school. In particular, individuals born in or before 1978 have OCP = 0; birth cohorts in or after 1979 have OCP = 1. Mother's age at birth is measured at individual i's birth. Log per capita household income is measure in 2011 values. Standard deviations are presented in parentheses.

Table 3.3. Descriptive Statistics: By One Child Policy Variable (OCPRelax) and Sex, Rural Area

Variable	Summary Statistics					
	OCPRelax = 0			OCPRelax = 1		
	All	Female	Male	All	Female	Male
School	8.87	8.28	9.35	10.1	10.1	10.1
	(2.74)	(2.86)	(2.54)	(2.48)	(2.52)	(2.45)
Sibsize	3.37	3.51	3.26	2.57	2.80	2.41
	(1.47)	(1.53)	(1.41)	(1.14)	(1.16)	(1.09)
Mother's Education	3.41	3.10	3.66	6.40	5.96	6.70
	(3.33)	(3.34)	(3.30)	(3.79)	(3.95)	(3.65)
Mother's Age at Birth	26.7	27.2	26.4	25.9	25.8	26.0
	(5.10)	(5.26)	(4.94)	(4.31)	(4.15)	(4.43)
Log Per Capita Household Income	8.23	7.99	8.42	8.71	8.52	8.83
	(0.97)	(0.89)	(0.99)	(1.10)	(1.03)	(1.13)
Age	27.2	23.3	30.3	22.8	21.6	23.6
	(6.86)	(4.22)	(6.98)	(3.42)	(2.79)	(3.59)
First-Born Sex (= 1 if female)	0.47	0.71	0.28	0.50	0.75	0.34
	(0.50)	(0.46)	(0.45)	(0.50)	(0.44)	(0.47)
N	1150	512	638	1008	411	597

Notes: The China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data are used. The rural-area sample includes individuals who are at least 18 years old and currently not in school. In particular, individuals born in or before 1978 have OCPRelax = 0; birth cohorts in or after 1979 have OCPRelax = 1. Mother's age is measured at individual i's birth. Log per capita is measure in 2011 values. Standard deviations are presented in parentheses.

Table 3.4. Descriptive Statistics: By First-Born Sex For Individuals with OCPRelax = 1

Variable	Summary Statistics			
	First-Born Sex = 1 (Female)		First-Born Sex = 0 (Male)	
	Mean	Standard Deviation	Mean	Standard Deviation
School	10.2	2.49	10.1	2.47
Sibsize	2.83	1.10	2.30	1.11
Mother's Education	6.32	3.76	6.47	3.83
Mother's Age at Birth	26.3	4.43	25.6	4.17
Log Per Capita Household Income	8.71	1.04	8.70	1.16
Age	22.5	3.24	23.1	3.58
N	508	508	500	500

Notes: The China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data are used. The rural-area sub-sample includes individuals who are at least 18 years old, currently not in school, and affected by the OCP and its subsequent relaxation (OCPRelax = 1). Mother's age is measured at individual i's birth. Log per capita is measure in 2011 values.

Table 3.5. Ordinary Least Square (OLS) and Two Stage Least Square (2SLS) Estimates of the Trade-off Effect between Quantity and Quality of Children: Urban Households

Variable	All	Female[1]	Male[1]	All	Female	Male
	(1)	(2)	(3)	(4)	(5)	(6)
	A. OLS Estimate: Schooling					
Sibsize	-0.44^{***}	-0.38^{***}	-0.53^{***}	-0.47^{***}	-0.38^{***}	-0.57^{***}
	(0.08)	(0.11)	(0.11)	(0.082)	(0.11)	(0.11)
	2SLS Estimate, Second Stage: Schooling					
Sibsize	-1.07^{***}	-2.57	-0.90^{***}	-1.20^{***}	-2.81^{*}	-0.91^{***}
	(0.38)	(1.91)	(0.33)	(0.35)	(1.60)	(0.32)
	2SLS Estimate, First Stage: Sibsize					
OCP	-0.53^{***}	-0.22	-0.73^{***}	-0.57^{***}	-0.27^{**}	-0.75^{***}
	(0.080)	(0.14)	(0.095)	(0.080)	(0.14)	(0.095)
Fixed Survey-Wave Effect	Y	Y	Y	N	N	N
Linear time trend	N	N	N	Y	Y	Y
Fixed Province Effect	Y	Y	Y	Y	Y	Y
N	951	436	515	951	436	515

Notes: The China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data are used. Covariates include: mother's education, mother's age at birth, sex, and log per capita household income. Robust standard errors are presented in parentheses. Individuals born in or before 1978 have OCP = 0 and individuals born in or after 1979 have OCP = 1. $***$ indicates significance at 1% level; $**$ indicates significance at 5% level; $*$ indicates significance at 10% level.

[1] Given the sample size, it is not of sufficient rank to perform model tests for the OLS estimation with fixed survey-wave effect and province effect using the female sub-sample and the male sub-sample, separately. Results should be interpreted with caution.

Table 3.6. Ordinary Least Square (OLS) and Two Stage Least Square (2SLS) Estimates of the Trade-off Effect between Quantity and Quality of Children: Rural Households

Variable	All	Female[1]	Male	All	Female[1]	Male
	(1)	(2)	(3)	(4)	(5)	(6)
	A. OLS Estimate: Schooling					
Sibsize	-0.12^{***}	-0.19^{***}	-0.077	-0.14^{***}	-0.20^{***}	-0.093
	(0.05)	(0.072)	(0.062)	(0.047)	(0.0.072)	(0.063)
	2SLS Estimate, Second Stage: Schooling					
Sibsize	0.16	0.32	0.0072	-0.0042	0.17	0.041
	(0.25)	(0.50)	(0.27)	(0.25)	(0.51)	(0.27)
	2SLS Estimate, First Stage: Sibsize					
OCPRelax	-0.60^{***}	-0.70^{***}	-0.53^{***}	-0.60^{***}	-0.64^{***}	-0.54^{***}
	(0.077)	(0.19)	(0.085)	(0.075)	(0.18)	(0.083)
OCPRelax × FB_sex	0.35^{***}	0.55^{***}	-0.060	0.33^{***}	0.53^{***}	-0.041
	(0.10)	(0.19)	(0.14)	(0.10)	(0.18)	(0.14)
Fixed Survey-Wave Effect	Y	Y	Y	N	N	N
Linear time trend	N	N	N	Y	Y	Y
Fixed Province Effect	Y	Y	Y	Y	Y	Y
N	2158	923	1235	2158	923	1235

Note: The China Health and Nutrition Survey (CHNS) 1989, 1991, 1993, 1997, 2000, 2004, 2006, 2009, and 2011 Data are used. Covariates include: mother's education, mother's age at birth, sex, first-born sex, and log per capita household income. Robust standard errors are presented in parentheses. Individuals born in or before 1978 have OCPRelax = 0 and individuals born in or after 1979 have OCPRelax = 1. $***$ indicates significance at 1% level; $**$ indicates significance at 5% level; $*$ indicates significance at 10% level.

[1] Given the sample size, it is not of sufficient rank to perform model tests for the OLS estimations using the female sub-sample. Results should be interpreted with caution.

Bibliography

John Abowd and Lars Vilhuber. The Sensitive of Economic Statistics to Coding Errors in Personal Identifiers. *Journal of Business and Economics Statistics*, 23(2):133–152, April 2005. JBES Joint Statistical Meetings invited paper with discussion and Rejoinder (April 2005):162–165.

John Abowd and Lars Vilhuber. National Estimates of Gross Employment and Job Flows from the Quarterly Workforce Indicators with Demographic and Industry Detail. *Journal of Econometrics*, pages 82–99, 2011. URL doi:10.1016/j.jeconom.2010.09.008.

John Abowd, John Haltiwanger, and Julia Lane. Integrated Longitudinal Employer-Employee Data for the United States. *The American Economic Review*, 94(2):224–229, 2004.

John Abowd, Bryce Stephens, Lars Vilhuber, Fredrik Andersson, Kevin McKinney, Marc Roemer, and Simon Woodcock. The LEHD Infrastructure Files and the Creation of the Quarterly Workforce Indicators. In Timothy Dunne, J. Bradford Jensen, and Mark Roberts, editors, *Producer Dynamics: New Evidence from Micro Data*, pages 149–230. University of Chicago Press, January 2009. URL http://www.nber.org/chapters/c0485.

Randy Albelda. Occupational Segregation by Race and Gender, 1958-1981. *Industrial and Labor Relations Review*, 39(3):404–411, 1986.

Joseph Altonji and Rebecca Blank. Race and Gender in the Labor Market. In Orley Ashenfelter and David Card, editors, *Handbook of Labor Economics, Vol. 3*, pages 3143–3259. The Netherlands: North-Holland Elsevier Science, 1999.

Elizabeth Ananat. The Wrong Side(s) of the Tracks: The Causal Effect of Racial Segregation on Urban Poverty and Inequality. *American Economic Journal: Applied Economics*, 3 (2):34–66, 2011. URL http://www.npc.umich.edu/news/events/econshocks/Ananat_Railroads_and_Segregation_21_Sept_2008.pdf.

Fredrik Andersson, Mónica García-Pérez, John Haltiwanger, Kristin McCue, and Seth Sanders. Workplace Concentration of Immigrants, November 2010. URL http://www.nber.org/papers/w16544. National Bureau of Economic Research Working Paper No. 16544.

Joshua Angrist and William Evans. Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size. *The American Economic Review*, 88(3):450–477, 1998.

Joshua Angrist and Jőrn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 2008.

Joshua Angrist, Victor Lavy, and Analia Schlosser. New Evidence on the Causal Link Between the Quantity and Quality of Children, December 2005. URL `http://www.nber.org/papers/w11835`. National Bureau of Economic Research Working Paper No. 11835.

Joshua Angrist, Victor Lavy, and Analia Schlosser. Multiple Experiments for the Causal Link between the Quantity and Quality of Children. *Journal of Labor Economics*, 28(4): 773–824, 2010.

Kenneth Arrow. The Theory of Discrimination. In Orley Ashenfelter and Albert Rees, editors, *Discrimination in Labor Markets*, pages 3–33. Princeton, NJ: Princeton University Press, 1973. URL `http://www2.econ.iastate.edu/classes/econ321/rosburg/Arrow%20-%20The%20Theory%20of%20Discrimination.pdf`.

Robert Barro and Jong-Wha Lee. A New Data Set of Educational Attainment in the World, 1950-2010. *Journal of Development Economics*, 104(C):184–198, 2013.

Gary Becker. *The Economics of Discrimination*. Chicago: University of Chicago Press, 1971.

Gary Becker and H. Gregg Lewis. On the Interaction between the Quantity and Quality of Children. *The Journal of Political Economy*, 81(2):S279–S288, 1973.

Gary Becker and H. Gregg Lewis. Interaction between the Quantity and Quality of Children. In Theodore Schultz, editor, *Economics of the Family: Marriage, Children, and Human Capital*, pages 81–90. UMI, 1974. URL `http://www.nber.org/chapters/c2963.pdf`.

Gary Becker and Nigel Tomes. Child Endowments and the Quantity and Quality of Children. *Journal of Political Economy*, 84(4):S143–S162, 1976.

Barbara Bergmann. Occupational Segregation, Wages and Profits When Employers Discriminate by Race or Sex. *Eastern Economic Journal*, 1(2):103–110, 1974.

Sandra Black, Paul Devereux, and Kjell Salvanes. The More the Merrier? The Effect of Family Size and Birth Order on Children's Education. *The Quarterly Journal of Economics*, 120(2):669–700, 2005.

Sandra Black, Paul Devereux, and Kjell Salvanes. Older and Wise? Birth Order and IQ of Young Men, August 2007. URL `http://ftp.iza.org/dp3007.pdf`. Institute for the Study of Labor (IZA) Discussion Paper No. 3007.

Judith Blake. Family Size and the Quality of Children. *Demography*, 18(4):421–442, 1981.

Francine Blau. *Equal Pay in the Office*. 1977.

Francine Blau, Marianne Ferber, and Anne Winkler. *The Economics of Women, Men, and Work*. Upper Saddle River, NY: Prentice Hall, 6 edition, 2010.

Kristin Butcher and Anne Case. The Effect of Sibling Sex Composition on Women's Education and Earnings. *The Quarterly Journal of Economics*, 109(3):531–563, 1994.

Gregorio Caetano and Vikram Maheshri. School Segregation and the Identification of Tipping Behavior, May 2014. URL `http://www.gregoriocaetano.net/Gregorio_Caetano/Gregorio_Caetano_files/caetano_maheshri_tipping.pdf`. Working Paper.

David Card, Alexandre Mas, and Jesse Rothstein. Tipping and the Dynamics of Segregation. *The Quarterly Journal of Economics*, 123(1):177–218, 2008a.

David Card, Alexandre Mas, and Jesse Rothstein. Are Mixed Neighborhoods Always Unstable? Two-sided And One-sided Tipping, November 2008b. URL `http://www.nber.org/papers/w14470`. National Bureau of Economic Research Working Paper No. 14470.

William Carrington and Kenneth Troske. On Measuring Segregation in Samples with Small Units. *Journal of Business & Economic Statistics*, 15(4):402–409, 1997.

William Carrington and Kenneth Troske. Interfirm Segregation and the Black/White Wage Gap. *Journal of Labor Economics*, 16(2):231–260, 1998.

U.S. Census Bureau. Business Dynamics Statistics (BDS), November 2013a. URL `https://www.census.gov/ces/dataproducts/bds/`.

U.S. Census Bureau. Longitudinal Business Database (LBD), July 2013b. URL `https://www.census.gov/ces/dataproducts/datasets/lbd.html`.

Hyowook Chiang, Kristin Sandusky, and Lars Vilhuber. LEHD Business Register Bridge Technical Documentation. *Longitudinal Employer – Household Dynamics Internal Document No. IP–LEHD–BRB–1.1.10*, June 2005. URL `www2.vrdc.cornell.edu/news/media/3/20050803-brb_ces_master.pdf`.

Dalton Conley. Sibship Sex Composition: Effects on Educational Attainment. *Social Science Research*, 29(3):441–457, 2000.

Dalton Conley and Rebecca Glauber. Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from exogenous Variation in Fertility. *The Journal of Human Resources*, 41(4):722–737, 2006.

Elisabeth Croll, Delia Davin, and Penny Kane. *China's One-Child Family Policy*. London: Macmillan, 1985.

David Cutler, Edward Glaeser, and Jacob Vigdor. The Rise and Decline of the American Ghetto. *Journal of Political Economy*, 107(3):455–506, 1999.

William Easterly. Empirics of Strategic Interdependence: The Case of the Racial Tipping Point. *The B.E. Journal of Macroeconomics*, 9(1), 2009.

Lena Edlund, Hongbin Li, Junjian Yi, and Junsen Zhang. Sex Ratios and Crime: Evidence from China's One Child Policy, December 2007. URL `http://ftp.iza.org/dp3214.pdf`. Institute for the Study of Labor (IZA) Discussion Paper No. 3214.

Roberto Fernandez and Isabel Ferdandez-Mateo. Networks, Race, and Hiring. *American Sociological Review*, 71(1):42–71, 2006.

Paul Frijters, Michael Shields, Nikolaos Theodoropoulos, and Stephen Price. Testing for Employee Discrimination Using Matched Employer-Employee Data: Theory and Evidence, June 2003. URL `http://ftp.iza.org/dp807.pdf`. Institute for the Study of Labor (IZA) Discussion Paper No. 807.

Claudia Goldin. A Pollution Theory of Discrimination: Male and Female Differences in Occupations and Earnings, June 2002. URL `http://www.nber.org/papers/w8985`. National Bureau of Economic Research Working Paper No. 8985.

Carlos Gradín, Coral del Río, and Olga Alonso-Villar. Occupational Segregation by Race and Ethnicity in the US: Differences Across States, 2011. URL `http://www.ecineq.org/milano/WP/ECINEQ2011-190.pdf`. Society for the Study of Economic Inequality Working Paper ECINEQ WP 2011-190, Universidade de Vigo.

Susan Greenlaugh. Shifts in China's Population Policy, 1984-1986: Views from the Central, Provincial, and Local Levels. *Population and Development Review*, 12(3):493–515, 1986.

Guang Guo and Leah VanWey. Sibship Size and Intellectual Development: Is the Relationship Causal? *American Sociological Review*, 64(2):169–187, 1999.

John Haltiwanger, Henry Hyatt, Erika McEntarfer, Liliana Sousa, and Stephen Tibbets. Firm Age and Size in the Longitudinal Employer-Household Dynamics Data, March 2014. URL `http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2423452`. U.S. Census Bureau Center for Economic Studies Paper No. CES-WP-14-16.

Emily Hannum. Educational Stratification by Ethnicity in China: Enrollment and Attainment in the Early Reform Years. *Demography*, 39(1):95–117, 2002.

Emily Hannum and Albert Park. *Education and Reform in China*. New York, NY: Routledge, 2007.

Emily Hannum, Jere Behrman, Meiyan Wang, and Jihong Liu. Education in the Reform Era. In Loren Brandt and Thomas Rawski, editors, *China's Great Economic Transformation*. Cambridge University Press, 2008. URL `http://works.bepress.com/emily_hannum/10`.

Bruce Hansen. Sample Splitting and Threshold Estimation. *Econometrica*, 68(3):575–603, 2000.

Eric Hanushek. The Trade-off between Child Quantity and Quality. *The Journal of Political Economy*, 100(1):84–117, 1992.

Robert Hauser and Hsiang-Hui Kuo. Does the Gender Composition of Sibships Affect Women's Educational Attainment? *The Journal of Human Resources*, 33(3):644–657, 1998.

Judith Hellerstein and David Neumark. Ethnicity, Language, and Workplace Segregation: Evidence from a New Matched Employer-Employee Data Set. *Annales d'Economie et de Statistique*, 71-72:19–78, July-December 2003.

Judith Hellerstein and David Neumark. Workplace Segregation in the United States: Race, Ethnicity, and Skill. *The Review of Economics and Statistics*, 90(3):459–477, 2008.

Judith Hellerstein, David Neumark, and Melissa McInerney. Changes in Workplace Segregation in the United States between 1990 and 2000: Evidence from Matched Employer-Employee Data. In Stefan Bender, Julia Lane, Kathryn Shaw, Fredrik Andersson, and Till von Wachter, editors, *The Analysis of Firms and Employees: Quantitative and Qualitative Approaches*, pages 163–195. National Bureau of Economic Research, Chicago, IL: The University of Chicago Press, 2008. URL `http://www.nber.org/chapters/c9115.pdf`.

Robert Higgs. Firm-Specific Evidence on Racial Wage Differentials and Workforce Segregation. *The American Economic Review*, 67(2):236–245, 1977.

Ron Jarmin and Javier Miranda. The Longitudinal Business Database, 2002. U.S. Census Bureau: CES Working Paper 02-17.

Robert Kaestner. Are Brothers Really Better? Sibling Sex Composition and Educational Attainment Revisited. *Journal of Human Resources*, 32(2):250–284, 1997.

Lisa Blau Kahn. The Long-term Labor Market Consequences of Graduating From College in A Bad Economy. *Labour Economics*, 17(2):303–316, 2010.

Lisa Blau Kahn and Erika McEntarfer. Worker Flows Over the Business Cycle: the Role of Firm Quality, 2013. URL `http://www.frbatlanta.org/documents/news/conferences/13employment_kahn.pdf?d=1&s=blogmb`. Working Paper.

Robert Kaufman. Assessing Alternative Perspectives on Race and Sex Employment Segregation. *American Sociological Review*, 67(4):547–572, 2002.

Daniel Kessler. Birth Order, Family Size, and Achievement: Family Structure and Wage Determination. *Journal of Labor Economics*, 9(4):413–426, 1991.

Mary King. Occupational Segregation by Race and Sex, 1940-88. *Monthly Labor Review*, 115(4):30–37.

John Knight and Lina Song. *The Rural-Urban Divide: Economic Disparities and Interactions in China*. Oxford: Oxford University Press, 1999.

David Lee and Thomas Lemieux. Regression Discontinuity Design in Economics. *Journal of Economic Literature*, 48(2):281–355, 2010.

Jungmin Lee. Sibling Size and Investment in Children's Education: An Asian Instrument. *Journal of Population Economics*, 21(4):855–875, 2008.

Ming-Hsuan Lee. The One Child Policy and Gender Equality in Education in China: Evidence from Household Data. *Journal of Family and Economic Issues*, 33:41–52, 2012.

Michael Leung and Junsen Zhang. Gender Preference, Biased Sex Ratio, and Parental Investments in Single-child Households. *Review of Economics of the Household*, 6(2): 91–110, 2008.

Roderick Little and Donald Rubin. *Statistical Analysis with Missing Data*. New York: John Wiley, 2 edition, 2002.

Charles Manski. Economic Analysis of Social Interactions. *The Journal of Economic Perspectives*, 14(3):115–136, 2000.

Marjorie McElory and Dennis Yang. Carrots and Sticks: Fertility Effects of China's Population Policies. *American Economic Review*, 90(2):389–392, 2000.

Kevin McKinney and Lars Vilhuber. LEHD-ECF Technical Documentation (Nonconfidential), 2005. URL `http://lehd.did.census.gov/led/library/tech_user_guides/ecf_ces_master.pdf`. U.S. Census Bureau: Internal Document No IP-LEHD-ECF-3.1.49.

Jessica Pan. Gender Segregation in Occupations: The Role of Tipping and Social Interactions, 2010. URL `http://www.aeaweb.org/aea/2011conference/program/retrieve.php?pdfid=270`. Working Paper.

Robert Parker, James Spletzer, and Michael Searson. The Business Establishment List - Standard Statistical Establishment List Comparison Project, June 2000. URL `www.bls.gov/ore/pdf/st010030.pdf`. Prepared for the Federal Economic Statistics Advisory Committee.

Edmund Phelps. The Statistical Theory of Racism and Sexism. *The American Economic Reveiw*, 62(4):659–661, 1972.

Nancy Qian. Missing Women and the Price of Tea in China: The Effect of Income on Sex Imbalance. *The Quarterly Journal of Economics*, 123(3):1251–1285, 2008.

Nancy Qian. Quantity-Quality and the One Child Policy: The Only-Child Disadvantage in School Enrollment in Rural China, May 2009. URL `http://www.nber.org/papers/w14973.pdf`. National Bureau of Economic Research Working Paper No. 14973.

Barbara Reskin, Debra McBrier, and Julie Kmec. The Determinants and Consequences of Workplace Sex and Race Composition. *Annual Review of Sociology*, 25:335–361, 1999.

Mark Rosenzweig and T. Paul Schultz. Fertility and Investment in Human Capital: Estimates of the Consequence of Imperfect Fertility Control in Malaysia. *Journal of Econometrics*, 36:163–184, 1987.

Mark Rosenzweig and Kenneth Wolpin. Testing the Quantity-Quality Fertility Model: The Use of Twins as a Natural Experiment. *Econometrica*, 48(1):227–240, 1980.

Mark Rosenzweig and Junsen Zhang. Do Population Control Policies Induce More Human Capital Investment? Twins, Birthweight, and China's One Child Policy. *Review of Economic Studies*, 76(3):1149–1174, 2009.

Eddie Salyers. An Assessment of Current Quality Assurance Practices and Ongoing Work to Develop a Comprehensive Quality Plan for U.S. Census Bureau Business Register, October 2004. URL `www.stats.gov.cn/english/18round/papers/200501/W020130912490883392336.doc`. Session 5, 18th International Roundtable on Business Survey Frames.

Thomas Schelling. Dynamic Models of Segregation. *Journal of Mathematical Sociology*, 1: 143–186, 1971.

Jesper Sorensen. The Organizational Demography of Racial Employment Segregation. *American Journal of Sociology*, 110(3):626–671, 2004.

Jeffrey Wooldridge. *Econometric Analysis of Cross-Section and Panel Data*. Cambridge, MA: MIT Press, 2 edition, 2010.

Yu Xie and Margaret Gough. Ethnic Enclaves and the Earnings of Immigrants, August 2009. URL `http://www.psc.isr.umich.edu/pubs/pdf/rr09-685.pdf`. Research Report, Population Studies Center, University of Michigan, Institute for Social Research.

Junfu Zhang. Tipping and Residential Segregation: A Unified Schelling Model. *Journal of Regional Science*, 51(1):167–193, 2011.

Junsen Zhang. Socioeconomic Determinants of Fertility in China. *Journal of Population Economics*, 3(2):105–123, 1994.

Junsen Zhang and Byron Spencer. Who Signs China's One-Child Certificate, and Why? *Journal of Population Economics*, 5(3):203–215, 1992.